VARIATIONAL ALGORITHMS FOR BAYESIAN INFERENCE IN LATENT GAUSSIAN MODELS

Een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde en Informatica

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 24 januari 2011
om 15:30 uur precies

door

Botond Cseke
geboren op 7 januari 1981
te Miercurea-Ciuc (Csíkszereda)

Promotor:
Prof. dr. T. M. Heskes

Manuscriptcommissie:
Prof. dr. H. J. Kappen
Prof. dr. M. Opper (Technische Universität, Berlin)
Dr. O. R. Zoeter (Xerox Research Center Europe, Grenoble)

SIKS Dissertation Series No. 2011-01

The research reported in this thesis has been carried out under the auspices of the Dutch Research School for Information and Knowledge Systems (SIKS), and the Institute for Computing and Information Sciences (ICIS) of the Radboud University Nijmegen.

The research has been funded by the Marie Curie EST program under the AI4IA project (Contract no. 514510) coordinated by SKF R&D (Nieuwegein, NL) and the Netherlands Organization for Scientific Research (NWO) under grant number 639.023.604.

ISBN: 978-90-9025750-1
Printing: Ipskamp Drukkers, Enschede.
Cover background: wordle.net, logo figurine: blackmancruz.com.

Contents

Chapter 1

A brief introduction to Bayesian models and variational inference

Graphical models are a powerful, graph theory based formalism to represent dependency relations in complex multivariate probabilistic models. They provide a unifying framework for addressing modeling and probabilistic inference problems in many areas of statistical and computational fields such as applied statistical analysis, statistical physics, bioinformatics, combinatorial optimization, signal processing and many more. In many of these areas, both the modeling and computational methods have long been formulated with the help of graphical models. The framework of probabilistic graphical models brings the above mentioned fields together and opens possibilities for formulating new models and algorithms.

Probabilistic graphical models are well-suited for formulating complex hierarchical Bayesian models where the conditional dependencies between the variables typically form a(n) (oriented) sparse interaction structure. This sparse (oriented) structure can be successfully exploited by *Markov chain Monte Carlo* (MCMC) sampling and *variational approximate inference* algorithms. The graphical model formalism, however, is not only a descriptive language since in many cases both the formulation and the characteristics of the inference algorithms, such as sufficient convergence properties, strongly depend on the underlying graph structure, that is, on the interaction structure between the variables. In this thesis, we apply the formalism and inference algorithms developed in the graphical models framework to solve problems in approximate Bayesian analysis.

1.1 Bayesian models

Bayesian inference is a statistical inference method that provides a principled way to update scientific hypotheses about certain quantities or the state of certain systems by using observational data. The hypotheses are typically expressed with the help of probabilities on the quantities or state variables under consideration. Let the variables in the vector $\boldsymbol{x}$ denote the quantities which we cannot observe directly and let the probability distribution or density $p(\boldsymbol{x}|M)$ express the uncertainty in $\boldsymbol{x}$ given our background information

expressed by the modeling choice M. When a (new) set of observations $\boldsymbol{y}$ becomes available, the information gained about $\boldsymbol{x}$ by observing $\boldsymbol{y}$ can then be used to update the uncertainty in $\boldsymbol{x}$. The assumed conditional probability density $p(\boldsymbol{y}|\boldsymbol{x}, M)$ expresses the uncertainty in $\boldsymbol{y}$ given a fixed $\boldsymbol{x}$ and the model choice M. In the following we consider both $\boldsymbol{x}$ and $\boldsymbol{y}$ continuous. The principled way to incorporate the information is by combining the probability densities using Bayes' rule (Cox, 1946). The "updated" uncertainty of $\boldsymbol{x}$ is then expressed by the conditional probability density $p(\boldsymbol{x}|\boldsymbol{y}, M)$ and computed according to

$$p(\boldsymbol{x}|\boldsymbol{y}, M) = \frac{p(\boldsymbol{y}|\boldsymbol{x}, M)p(\boldsymbol{x}|M)}{p(\boldsymbol{y}|M)} \qquad \text{(Bayes' rule)}.$$

The proportionality factor $p(\boldsymbol{y}|M)$ is independent of $\boldsymbol{x}$ and it expresses the probability of observing $\boldsymbol{y}$ given the model M. It is computed by averaging the probability of observing $\boldsymbol{y}$ over all possible values of $\boldsymbol{x}$, that is

$$p(\boldsymbol{y}|M) = \int d\boldsymbol{x}\, p(\boldsymbol{x}|M)p(\boldsymbol{y}|\boldsymbol{x}, M) \qquad \text{(evidence)}.$$

This quantity is called the *evidence* and can be used to compare two different modeling choices M_1 and M_2. The probability density $p(\boldsymbol{x}|M)$ is called the *prior* while $p(\boldsymbol{x}|\boldsymbol{y}, M)$ is called the *posterior* probability density. A crucial property of the Bayesian updating rule is that the information gained by a set of observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ can be "added" incrementally to update the uncertainty in $\boldsymbol{x}$, that is,

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}_1, M) &\propto p(\boldsymbol{x}|M)p(\boldsymbol{y}_1|\boldsymbol{x}, M), \\ p(\boldsymbol{x}|\boldsymbol{y}_1, \boldsymbol{y}_2, M) &\propto p(\boldsymbol{x}|\boldsymbol{y}_1, M)p(\boldsymbol{y}_2|\boldsymbol{x}, M), \ \ldots \end{aligned}$$

This allows us to update the probability density $p(\boldsymbol{x}|M)$, whenever any information in terms of observations $\boldsymbol{y}_t, t = 1, \ldots, T$ becomes available. Besides updating the uncertainties and assessing the probability of certain observations, the other important task of Bayesian inference is to provide predictions for future observations. That is, given the *prior* probability density $p(\boldsymbol{x}|M)$ and the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, the task is to compute the probability density $p(\boldsymbol{y}_*|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T, M)$ of a new observation $\boldsymbol{y}_*$. This probability density follows from the application of Bayes' rule.

Up to this point we assumed that $\boldsymbol{x}$ and $\boldsymbol{y}$ denote sets of continuous variables. Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ and let $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^m$ denote all the considered observable variables. When applying Bayesian inference we are typically interested in the following quantities:

(1) the evidence $p(\boldsymbol{y}|M)$ given the observed variables $\boldsymbol{y}$,

(2) the posterior marginal probability densities $p(\boldsymbol{x}_I|\boldsymbol{y}, M) = \int d\boldsymbol{x}_{\setminus I}\, p(\boldsymbol{x}|\boldsymbol{y}, M)$ or certain posterior expectations $\mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y},M)}[f(\boldsymbol{x}_I)]$ of a set of variables x_I with $I \subset \{1, \ldots, n\}$,

(3) the predictive probability densities $p(\boldsymbol{y}_*|\boldsymbol{y}, M) = \int d\boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y}, M)p(\boldsymbol{y}_*|\boldsymbol{x}, M)$,

(4) the posterior conditional probability densities $p(\boldsymbol{x}_{I_1}|\boldsymbol{x}_{I_2}, \boldsymbol{y}, M)$ of two disjoint sets

of variables $\boldsymbol{x}_{I_1}$ and $\boldsymbol{x}_{I_2}$,

(5) the mode $\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}, M)$ of the posterior density (this quantity is particularly important in cases when the x_is are discrete variables).

In this thesis we focus on approximating the posterior marginal probability densities. Except for a relatively restricted class of prior and likelihood models the integrals or summations required to compute the above listed quantities are intractable. The complexity of the standard numerical quadratures typically increases exponentially with the number of variables n, therefore models with more than, say, ten variables are not tractable with these methods.

A typical way to go around the problem of intractability is to approximate the posterior density with a density for which the integrals are tractable. *Sampling methods* (e.g. Metropolis et al., 1953; Casella and George, 1992; Neal, 1993; Rue and Held, 2005) achieve this by representing or approximating the posterior probability density by a sum of delta functions

$$p(\boldsymbol{x}|\boldsymbol{y}, M) = \frac{1}{S}\sum_{s=1}^{S} \delta_0(\boldsymbol{x} - \boldsymbol{x}_s)$$

at locations $\boldsymbol{x}_s$ sampled independently from the posterior density. There is a vast literature on sampling methods. Until the relatively recent spread of the variational methods, approximate inference was almost exclusively dominated by them. The crux in sampling is to find a set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_S\}$ of reasonable cardinality such that the samples are statistically independent and they cover the areas where the main mass of the posterior density lies. There are various methods that yield good coverage of the main mass and account for the dependence of the samples or try to minimize dependence. We refer the reader to Neal (1993) and references therein. Sampling methods are typically slow compared to the other well known classes of approximate inference methods, but they come with theoretical guarantees for convergence in the limit of infinite sample size. Therefore, they are used as a golden standard to compare the accuracy of other methods to, and because of this, they are indispensable in any approximate inference setting.

1.2 Probabilistic graphical models

The Bayesian framework treats all model quantities equally, that is, the variables in $\boldsymbol{x}$ can be latent variables, model parameters and nuisance parameters as well. Remember that the observations $\boldsymbol{y}$ are also treated as variables. In the majority of cases, the variables in $\boldsymbol{x}$ and $\boldsymbol{y}$ have a complex hierarchical dependency structure and both the prior $p(\boldsymbol{x}|M)$ and the likelihood $p(\boldsymbol{y}|\boldsymbol{x}, M)$ are defined in terms of conditional dependencies

$$p(\boldsymbol{x}|M) = \prod_j p(x_j|\boldsymbol{x}_{\pi_x(j)}), \tag{1.1}$$

$$p(\boldsymbol{y}|\boldsymbol{x}, M) = \prod_i p(y_i|\boldsymbol{x}_{\pi_y(i)}), \tag{1.2}$$

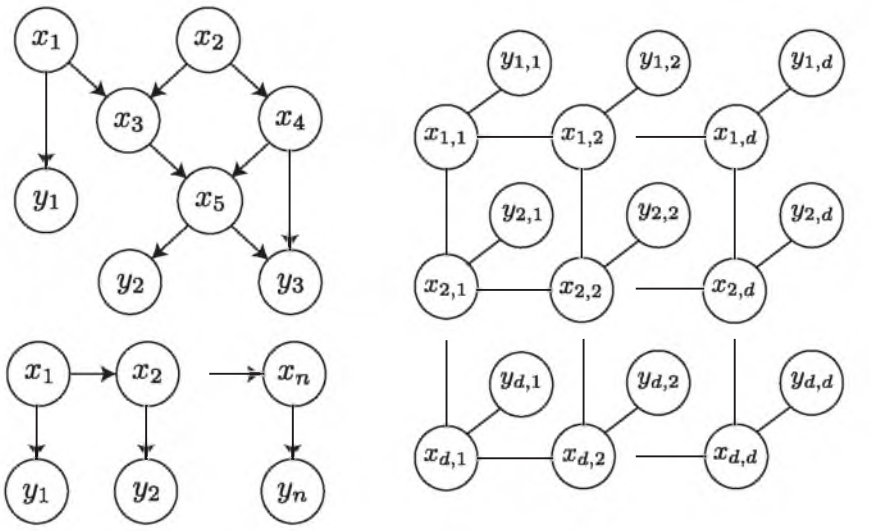

Figure 1.1: Graphical representations of an arbitrary Bayesian network (top-left), a hidden Markov model (bottom-left), and a probabilistic model based on a two dimensional Ising model (right), where the pairwise interaction are represented by undirected edges.

where $\pi_x(j)$ and $\pi_y(i)$ denote the index set of variables on which x_j and y_i depend. Examples include the linear dynamical model or the hidden Markov model where $p(\boldsymbol{x}|M) = p(x_1|M)\prod_j p(x_j|x_{j-1}, M)$ and $p(\boldsymbol{y}|\boldsymbol{x}, M) = \prod_i p(y_i|x_i, M)$ or the hierarchical regression models like the ones presented in Chapter 4. The factorization properties of the prior and the likelihood yield great computational advantages when sampling methods like Gibbs sampling (Geman and Geman, 1984; Casella and George, 1992) or more generally Markov Chain Monte Carlo sampling (e.g. Neal, 1993) are applied. For this reason, it is often desirable to formulate the models in terms of dependency structures and to design inference algorithms that exploit these structures.

Bayesian networks

The dependency structure defined by equations (1.1) and (1.2) can often be represented by directed acyclic graphical structures called *Bayesian networks* (e.g. Cowell et al., 1999). Let $G = (V, E)$ be a directed acyclic graph such that the nodes in V correspond to the variables x_j and y_i. We label the vertices by the indices of variables x_j and y_i they correspond to. The factors in equations (1.1) and (1.2) can be represented by adding directed edges $(k, j) \in E$ for all $k \in \pi_x(j)$ and $(k, j) \in E$ for all $k \in \pi_y(i)$ accordingly. The resulting graph faithfully represents the conditional relations between the variables of the joint model $p(\boldsymbol{y}|\boldsymbol{x}, M)p(\boldsymbol{x}|M)$ and it provides a generative model or sampling distribution for all the involved variables.

When the computation involving the posterior density $p(\boldsymbol{x}|\boldsymbol{y}, M)$ or its mareginals are not analytically tractable, we have to resort to numerical approximation methods. This re-

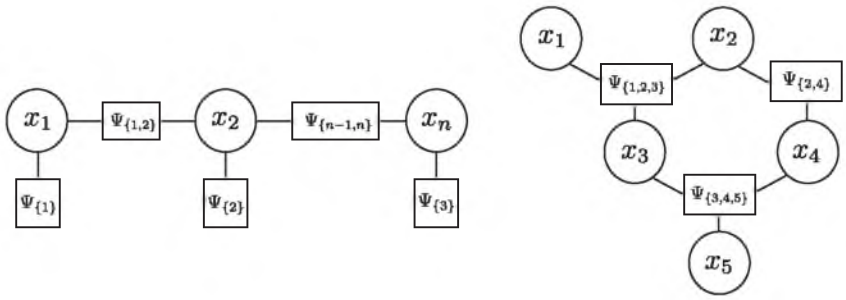

Figure 1.2: Factor graphs corresponding to the Bayesian networks in Figure 1.1.

quires fixing $\boldsymbol{y}$ to the observed values and approximating the posterior with a parametric distribution or sampling from it. Setting $\boldsymbol{y}$ to a fixed value alters the dependency structure, because it introduces interactions between variables that no longer represent faithful conditional dependencies. However, the posterior density still factorizes over a product of functions that depend on the variable sets $\{j\} \cup \pi_x(j)$ and $\pi_y(i)$. It is this factorization property that most sampling and variational approximate inference algorithms can make use of.

The conditional dependencies between variables can also be represented by undirected graphs and all Bayesian networks can be turned into undirected graphical models (e.g. Lauritzen, 1996; Cowell et al., 1999), like the ones in Figure 1.1 and 1.2. Since we are only interested in inference, we focus on a different graphical representation called *factor graphs*.

Factor graphs

Not all dependency relations between variables can be represented by conditional densities. As mentioned above, from the point of view of inference it is not the conditional dependency structure of the joint $p(\boldsymbol{y}, \boldsymbol{x}|M)$ what matters most, but the factorization of the posterior density. The most suitable tool to describe the factorization is the factor graph formalism (e.g. Kschischang et al., 2001). Let the posterior density factorize as

$$p(\boldsymbol{x}|\boldsymbol{y}, M) \propto \prod_{\alpha} \Psi_{\alpha}(\boldsymbol{x}_{\alpha})$$

where $\alpha \subseteq \{1, \ldots, n\}$ denotes the set of indices the factor Ψ_{α} depends on. The factor graph is defined as a bi-partite graph: the vertices are partitioned into two sets, the set of factors indexed by the sets α and the set of nodes $x_1, \ldots, x_n$; the undirected edges are between factors and variables only, namely, a factor Ψ_{α} is connected to all variables x_j with $j \in \alpha$. The definition of the α sets may be completely arbitrary, but it can also strongly depend on the algorithm at hand. For example, in the above mentioned case of hidden Markov models it is suitable to choose $\alpha = \{i, i-1\}$ and $\alpha = \{k\}$ such that $\Psi_{\{i,i-1\}}(x_i, x_{i-1}) = p(x_i|x_{i-1}, M)$ and $\Psi_{\{k\}}(x_k) = p(y_k|x_k)$ respectively (see Figure 1.2). A typical example where the prior is not defined in terms of conditional

distributions is the Ising model, where $p(\boldsymbol{x}|\boldsymbol{Q})$ is the maximum entropy distribution of a magnetic spin system with a fixed average energy. In this case, the variables $x_i \in \{-1, 1\}$ represent the magnetic spins and $-\log p(\boldsymbol{x}|\boldsymbol{Q}) = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \log Z(\boldsymbol{Q})$, where $\boldsymbol{Q}$ is an $n \times n$ matrix specifying the interactions between the variables and $Z(\boldsymbol{Q})$ is the normalization constant. The variables $\boldsymbol{x}$ are typically organized in a spatial grid structure, implying a sparse $\boldsymbol{Q}$. In many applications the likelihood factorizes as $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_i p(y_i|x_i)$, therefore, we can define the factors of this model as $\Psi_{\{i,j\}}(x_i, x_j) = \exp(-Q_{ij}x_ix_j)$ for all i and j with $Q_{ij} \neq 0$ and $\Psi_{\{k\}}(x_k) = p(y_k|x_k)$. This model can be represented by the undirected graph shown in Figure 1.1.

Using the factor graph formalism, it is possible to do exact computation of the marginals. When the factor graph has a tree structure and the variables are all continuous or all discrete, one can schedule the integration or summation such that the marginals are correct up to numerical roundoff errors or numerical quadrature rules. This algorithm is called *message passing* or *belief propagation* (e.g. Lauritzen, 1996). It is a dynamic programming scheme that uses the properties of the summation and integration operators and the structure of the factor graph to simplify the marginalization to a sequence of sums or integrals on low dimensional factors. The *forward-backward* algorithm for hidden Markov models or the *filtering* and *smoothing* computations for linear dynamical systems (e.g. Minka, 1998) are instances of the message passing algorithm. In models with both discrete and continuous variables, like switching linear systems (e.g. Zoeter and Heskes, 2005a), the algorithm can become intractable. In factor graphs with loops, after a suitable reorganization of the factors, one can define a tree structured factor graph on groups of variables called the *junction tree* and run the message passing algorithm on it. However, the junction tree construction often results in factors that connect large numbers of variables and thus even local computations can be computationally demanding. Murphy et al. (1999) pointed out that the message passing algorithm can yield relatively good approximations of the marginals even when it is run on loopy factor graphs. Thus, in situations when the message passing on the junction tree is too costly or unfeasible, one can opt for the (loopy) message passing algorithm.

1.3 Variational approximate inference methods

A different approach to approximate the marginals is to formulate the marginalization as an optimization problem. This requires the definition of an error function and a set of candidate optimizers. The Kullback-Leibler divergence $\mathrm{D}\,[q\,||\,p] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log[q(\boldsymbol{x})/p(\boldsymbol{x})]$ or $\mathrm{D}\,[q\,||\,p] = \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log[q(\boldsymbol{x})/p(\boldsymbol{x})]$ and its generalizations are popular quantities to measure the difference between two probability distributions or density functions, therefore, they are often used as objective functions when approximating probability densities.

Variational approximations

Assuming that the candidate distributions q factorize as $q(\boldsymbol{x}) = \prod_j q_j(x_j)$, the minimization of $\mathrm{D}[p||\prod_j q_j]$ would yield the optimal (exact) solution $q_j(x_j) = \int d\boldsymbol{x}_{\setminus j}\, p(\boldsymbol{x})$, but it would also involve the direct computation of $\int d\boldsymbol{x}_{\setminus j}\, p(\boldsymbol{x})$. A way to relax this problem is to change the roles of q and p in $\mathrm{D}\,[q\,||\,p]$ and minimize $\mathrm{D}[\prod_j q_j||p]$ under the constraints

that the q_js are positive and normalize to one. The divergence $\mathrm{D}\left[q \,||\, p\right]$ can be written as the sum of the so-called negative energy term $-\mathrm{E}_q[\log p]$ and the negative entropy $\mathrm{E}_q[\log q]$ which in this case decomposes as $\mathrm{E}_q[\log q] = \sum_j \mathrm{E}_{q_j}[\log q_j]$. Although the objective is not jointly convex in $q_1, \ldots, q_n$, one can try to find a local minimum by running the fixed point iteration derived from the stationary conditions of the corresponding Lagrangian (Jaakkola, 2000). The fixed point iteration has similar computational advantages as the message passing algorithm, when $p(\boldsymbol{x})$ factorizes into a product of terms, in order to update, q_j, one only has to compute expectations of the log-factors which are depending on x_j. When convergence occurs, one can use the q_js as approximate marginals of $p(\boldsymbol{x})$. The minimum of the objective gives an approximation of the negative log normalization constant of p. This corresponds to the negative log evidence when p is the posterior density $p(\boldsymbol{x}|\boldsymbol{y}, M)$ in a Bayesian model. This method is referred to as the *mean field* approximation.

In case of continuous variables one can also define the candidate probability densities q as belonging to a family of parametric multivariate densities and try to minimize $\mathrm{D}\left[q \,||\, p\right]$. This can be tractable because it requires computing the expectations of the log-factors and the entropy of q (e.g. Opper and Archambeau, 2009).

Variational approximations in graphical models

The tractability of the mean field approximation is due to the decomposition of the entropy which follows from the factorization assumptions in q. Let us consider a model with discrete variables such that $p(\boldsymbol{x}) \propto \prod_{ij} \Psi_{ij}(x_i, x_j)$ has a tree structured factor graph. In this case, one can go beyond the factorized approximation and compute the exact marginals $p(x_k)$ and $p(x_i, x_j)$. The procedure works as follows. Let us assume that the distribution q has the form

$$q(\boldsymbol{x}) = \prod_k q_k(x_k) \prod_{i \sim j} \frac{q_{ij}(x_i, x_j)}{q_i(x_i) q_j(x_j)} \quad \text{with} \quad \sum_{x_j} q_{ij}(x_i, x_j) = q_i(x_i) \tag{1.3}$$

and the set of ijs follows the same (tree) structure as p. Then $q_{ij}(x_i, x_j)$ and $q_k(x_k)$ are exact marginals of q and the entropy can be computed exactly in terms of q_k and q_{ij}. The objective $\mathrm{D}\left[q \,||\, p\right]$ is jointly convex in q_k and q_{ij} and the constraints are linear. The iteration derived from the stationary points of the Lagrangian can be rewritten to have the same form as the message passing algorithm (e.g. Yedidia et al., 2000). When the factorizations in p and q do not lead to a tree structured factor graph, the q_ks and q_{ij}s are not marginals of q and the entropy cannot be computed exactly. In this case, $\mathrm{D}\left[q \,||\, p\right]$ can be approximated by using the so-called Bethe approximation of the negative entropy

$$\mathrm{E}_q[\log q] \approx \sum_{i \sim j} \mathrm{E}_{q_{ij}}[\log q_{ij}] + \sum_k (1 - n_k) \mathrm{E}_{q_k}[\log q_k] \,,$$

where n_k is the number of variables x_k is connected to through the factorization in $p(\boldsymbol{x})$. The resulting objective is called the *Bethe free energy*. The fixed point iteration can then be rewritten in the form of the loopy version of the message passing algorithm (e.g. Yedidia et al., 2000). The optimal q_ks and q_{ij}s are approximations of the correspond-

ing marginal probabilities and the minimum of the Bethe free energy can be used as an approximation of the negative log normalization constant (the log evidence when p is the posterior density of a Bayesian model). Although the objective has multiple local minima, Heskes (2003) and Watanabe and Fukumizu (2009) have shown that stable fixed points of the message passing algorithm are local minima of the Bethe free energy. In the last decade there has been an intensive research in this area. See Chapter 2 for a broader overview.

These approximations can also be adapted to models with continuous variables. In many important applications with discrete variables, the parameters of the messages in the message passing algorithm can be computed exactly. For continuous variables, this is not the case unless the Ψ_{ij}s belong to a family of functions that is closed under marginalization, like for example the Gaussian family discussed in Chapter 2. When the family is not closed, the local marginalization in the message passing algorithm can be replaced by operations involving projections using the Kullback-Leibler divergence. This leads to an algorithm referred to as expectation propagation (Minka, 2001, 2005). For further details about expectation propagation the reader is referred to Chapters 3 and 4 and references therein.

1.4 Thesis outline

The results in this thesis are based on applications of the expectation propagation algorithm to approximate marginals in models with Gaussian prior densities.

In Chapter 2, we start out with a model where the likelihood $p(\boldsymbol{y}|\boldsymbol{x}, M)$ is Gaussian as well and study the properties of the message passing algorithm and the corresponding Bethe free energy. It turns out that although in terms of functional parameters q_k and q_{ij} the free energy has the same property as in the discrete case, when expressing it in a parametric form that incorporates the marginal consistency constraints (as in (1.3)), its behavior is quite surprising. While in the discrete case the free energy is a bounded function, in the Gaussian case it can be unbounded when expressed in terms of the moment parameters of the approximate marginals q_k and q_{ij}. The typical relaxations applied in the discrete case (e.g. Wainwright et al., 2003; Wiegerinck and Heskes, 2003) seem to achieve the opposite effect by creating a convex objective with an unbounded global minimum. We show that the stable fixed points of the Gaussian message passing algorithm are local minima of the Gaussian free energy and that both the convergence of the message passing algorithm and the existence of local minima is more likely for relaxation parameters that move the free energy closer to the mean field free energy. We also give sufficient and necessary conditions for the boundedness of the Gaussian Bethe free energy.

In Chapter 3, we address the problem of approximating posterior marginals in models where $p(\boldsymbol{x}|M)$ is a Gaussian prior and the non-Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{x}, M)$ factorizes as $p(\boldsymbol{y}|\boldsymbol{x}, M) = \prod_i p(y_i|x_i, M)$. The methods we propose are not restricted to these models, but they are particularly well-suited for them. The approximate posterior marginals in these models are typically computed by approximating the non-Gaussian posterior density with a multivariate Gaussian density q either using the variational objective $\mathrm{D}\,[q\,||\,p]$ or by expectation propagation. The Gaussian marginals $q(x_i)$ are then used as approximations of the posterior marginals. In Chapter 3 we go beyond these Gaussian

marginal approximations and we derive a framework to improve on the Gaussian $q(x_i)$s. The improved marginals seem to perform well in the comfort zone of EP, that is, in models where the posterior density is log-concave. Although we do not provide an estimate or an upper bound on the error, the approximations have the nice property that they can be gradually improved whenever better accuracy is needed.

In Chapter 4, we define a multivariate scale mixture distribution that can be used as a sparsifying prior in the context of linear regression and logistic regression. We derive an efficient expectation propagation algorithm to do approximate inference in these models. We use these models to do approximate Bayesian inference for assessing the activation of brain areas in task-related MEG and fMRI experiments. The multivariate prior we introduce is based on the scale mixture representation of the univariate double exponential prior and it is defined with the aim to introduce prior correlations between the magnitudes of the regression coefficients. This was motivated by the observation that in many MEG and fMRI applications the activations have smooth spatial and temporal patterns, that is, neighboring brain areas (in space, in time, or both) are likely to have similar activation levels. The prior keeps the regression coefficients a priori uncorrelated, but it correlates their magnitudes. The symmetry properties of the prior lead to posterior densities that imply block diagonal correlation structures. The approximating multivariate Gaussians inherit this property. The block diagonal covariance structure and the typically underdetermined regression models make the computational complexity of EP to scale linearly with the number of regression coefficients. We show that the importance maps created from the approximate posterior moments of the scale parameters are meaningful and neuro-biologically reasonable.

Chapter 2

Bethe Free Energies and Message Passing in Gaussian Models

Summary

We address the problem of computing approximate marginals in Gaussian probabilistic models by using mean field and fractional Bethe approximations. As an extension of Welling and Teh (2001), we define the Gaussian fractional Bethe free energy in terms of the moment parameters of the approximate marginals and derive an upper and lower bound for it. We give necessary conditions for the Gaussian fractional Bethe free energies to be bounded from below. It turns out that the bounding condition is the same as the pairwise normalizability condition derived by Malioutov et al. (2006) as a sufficient condition for the convergence of the message passing algorithm. Using the same line of argument as Watanabe and Fukumizu (2009) used for loopy belief propagation in discrete models, we show that stable fixed points of the Gaussian message passing are local minima of the Gaussian Bethe free energy. By a counterexample, we disprove the conjecture in Welling and Teh (2001): even when the Bethe free energy is not bounded from below, it can still possess a local minimum to which the minimization algorithms can converge. The material presented in this chapter is an extension of Cseke and Heskes (2008)[1] and it contains the results reported in Cseke and Heskes (2010b).

2.1 Introduction

One of the major tasks of probabilistic inference is calculating marginal probabilities of a set of variables given some observations. In case of Gaussian models, the complexity of computing marginals might scales cubically with the number of variables, while for models with discrete variables it often leads to intractable computations. Computations can be made faster or tractable by using approximate inference methods like the mean field approximation (e.g. Jaakkola, 2000) and the Bethe approximation (e.g. Yedidia et al., 2000). These methods were developed mainly for discrete probabilistic graphical models,

[1] B. Cseke and T. Heskes, *Bounds on the Bethe free energy for Gaussian networks*, UAI-2008, pages 97–104.

but they are applicable to Gaussian models as well. However, there are important differences in their behavior for the discrete and Gaussian cases. For example, while in discrete models the error function of the Bethe approximation—called Bethe free energy—is bounded from below (Heskes, 2004; Watanabe and Fukumizu, 2009), in Gaussian models this might not always the case (Welling and Teh, 2001).

The study of the Bethe free energy of Gaussian models is also motivated by their importance for the study of conditional Gaussian models. Conditional Gaussian or hybrid graphical models, such as switching Kalman filters (e.g. Zoeter and Heskes, 2005a), combine both discrete and Gaussian variables. Approximate inference in these models can be carried out by expectation propagation (e.g. Minka, 2004, 2005). Expectation propagation can be viewed as a generalization of the Bethe approximation, where marginalization constraints are replaced by expectation constraints (e.g. Heskes et al., 2005). Therefore, studying the properties of the Bethe free energy can reveal some of the convergence properties of expectation propagation. In order to understand the properties of the Bethe free energy of hybrid models, a good understanding of the two special cases of discrete and Gaussian models is needed. While the properties of the Bethe free energy of discrete models have been studied extensively in the last decade and are well understood (e.g. Yedidia et al., 2000; Heskes, 2003; Wainwright et al., 2003; Watanabe and Fukumizu, 2009), the properties of the Gaussian Bethe free energy have been studied much less.

The message passing algorithm is a well established method for finding the stationary points of the Bethe free energy (e.g. Pearl, 1988; Yedidia et al., 2000; Heskes, 2003). It works by locally updating the approximate marginals and has been successfully applied in both discrete (e.g. Murphy et al., 1999; Wainwright et al., 2003) and Gaussian models (e.g. Weiss and Freeman, 2001; Rusmevichientong and Roy, 2001; Malioutov et al., 2006; Johnson et al., 2009a,b; Nishiyama and Watanabe, 2009; Bickson, 2009). Gaussian message passing is the simplest case of a free-energy based message passing algorithm on models with continuous variables, therefore, it is important to understand its behavior.

Gaussian message passing has many practical applications like in distributed averaging (e.g. Moallemi and Roy, 2006), peer-to-peer rating, linear detection, SVM regression (e.g. Bickson, 2009) and more generally in problems that involve solving large sparse linear systems or approximating the marginal variances of large sparse Gaussian systems typically encountered in distributed computing settings. For further applications the reader is referred to Bickson (2009) and references therein.

Finding sufficient conditions for the convergence of message passing in Gaussian models has been successfully addressed by many authors. Using the computation tree approach, Weiss and Freeman (2001) proved that message passing converges whenever the precision matrix—inverse covariance—of the probability distribution is diagonally dominant[2]. With the help of an analogy between message passing and walk–sum analysis, Malioutov et al. (2006) derived the stronger condition of pairwise normalizability[3]. A different approach was taken by Welling and Teh (2001), who directly minimized the Bethe free energy with regard to the parameters of approximate marginals, conjecturing that Gaussian message passing converges if and only if the free energy is bounded from

[2]The matrix $\boldsymbol{A}$ is diagonally dominant if $|A_{ii}| > \sum_{j \neq i} |A_{ij}|$ for all i.

[3]Following Malioutov et al. (2006), we call a Gaussian distribution pairwise normalizable if it can be factorized into a product of normalizable "pair" factors, that is, $p(x_1, \ldots, x_n) = \prod_{ij} \Psi_{ij}(x_i, x_j)$ such that all Ψ_{ij}s are normalizable.

below. Their experiments showed that message passing and direct minimization either converge to the same solution or both fail to converge. We take a similar approach, that is, instead of analyzing the properties of the Gaussian message passing algorithm, we choose to study the properties of the Gaussian Bethe free energy. This will help us to draw conclusions about the existence of local minima, the possible stable fixed points to which message passing can converge.

This chapter is structured as follows. In Section 2.2 we introduce Gaussian Markov random fields and the message passing algorithm. In Section 2.3 we define the Gaussian fractional Bethe free energies parameterized by the moment parameters of the approximate marginals and derive boundedness conditions for them. These two sections are based on our earlier chapter (Cseke and Heskes, 2008). In Section 2.4 we analyze the stability properties of the Gaussian message passing algorithm and, using a similar line of argument as Watanabe and Fukumizu (2009), we show that its stable fixed points are indeed local minima of the Bethe free energy. We conclude the chapter with a few experiments in Sections 2.5 and 2.6 supporting our results and their implications.

2.2 Approximating marginals in Gaussian models

The probability density of a Gaussian random vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined in terms of canonical parameters $\boldsymbol{h}$ and $\boldsymbol{Q}$ as

$$p(\boldsymbol{x}) \propto \exp\left\{\boldsymbol{h}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right\}, \tag{2.1}$$

where $\boldsymbol{Q}$ is s positive definite matrix. The expectation $\boldsymbol{m}$ and the covariance $\boldsymbol{V}$ of $\boldsymbol{x}$ is then given by $\boldsymbol{m} = \boldsymbol{Q}^{-1}\boldsymbol{h}$ and $\boldsymbol{V} = \boldsymbol{Q}^{-1}$ respectively. In many real world application the matrix $\boldsymbol{Q}$ is sparse with a typically very low density, that is, the number of elements in $\boldsymbol{Q}$ scales with the number of variables n.

This probability density can also be defined in terms of an undirected probabilistic graphical model commonly known as Gaussian Markov random field (GMRF). Since the interactions between the variables in p are pairwise, we can associate the variables x_i to the nodes $v \in V = \{1, \ldots, n\}$ of an undirected graph $G = (V, E)$, where the edges $e \in E \subseteq V \times V$ of the graph stand for the non-zero off-diagonal elements of $\boldsymbol{Q}$. The density p can then be defined as the product

$$p(\boldsymbol{x}) \propto \prod_{(i,j)\in E} \Psi_{ij}(x_i, x_j) \tag{2.2}$$

of Gaussian functions $\Psi_{ij}(x_i, x_j)$ (also called potentials) associated with the edges $e = (i, j)$ of the graph. If $\boldsymbol{h}$ and $\boldsymbol{Q}$ are given then we can define the potentials as

$$\Psi_{ij}(x_i, x_j) = \exp\left\{\gamma_{ij}^i h_i x_i + \gamma_{ij}^j h_j x_j - \gamma_{ij}^i Q_{ii} x_i^2/2 - \gamma_{ij}^j Q_{jj} x_j^2/2 - Q_{ij} x_i x_j\right\},$$

where $\sum_{i\sim j}\gamma_{ij}^i = 1$ and $\sum_{j\sim i}\gamma_{ij}^j = 1$ are partitioning $\boldsymbol{h}$ and $\boldsymbol{Q}$ into the respective factors. In practice, however, the factors Ψ_{ij} might be given by the problem at hand and $\boldsymbol{h}$ and $\boldsymbol{Q}$ as well as γ_{ij}^i and γ_{ij}^j computed by summing their parameters and computing

the partitioning respectively. Without loss of generality, we can and we will use $Q_{ii} = 1$, since the results in the chapter can be easily re-formulated for general $\boldsymbol{Q}$s by a rescaling of the variables.

Exact calculation all marginals, can be done by solving the linear system $\boldsymbol{m} = \boldsymbol{Q}^{-1}\boldsymbol{h}$ and performing a sparse Cholesky factorization $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{Q}$ followed by solving the Takahashi equations (Takahashi et al., 1973). A method for solving the Takahashi equations is presented in Section A.2 in the Appendix. The complexity of the latter two scales with $nnzeros\left(\boldsymbol{Q}\right)^2/n$.

An alternative option to calculate the marginal means and to *approximate marginal variances* is to run the Gaussian message passing algorithm in the probabilistic graphical model associated with the representation (2.2). Gaussian message passing is the Gaussian variant of message passing (Pearl, 1988), which is a dynamical programming algorithm introduced to compute marginal densities in discrete probabilistic models with pairwise interactions and tree-structured graphs G. However, it turned out that by running it in loops on graphs with cycles, it yields good approximations of the marginal distributions (Murphy et al., 1999). Weiss and Freeman (2001) showed that when the Gaussian (loopy) message passing is converging it computes the exact mean parameters $\boldsymbol{m}$, thus it can also be used for solving linear systems (e.g. Bickson, 2009). Message passing works by updating and passing directed messages along the edges of the graph which, in case the algorithm converges, are then used to compute (approximate) marginal probability distributions. The Gaussian and the discrete algorithms have the same functional form with the exception of the summation (discrete case) and integration operators (Gaussian case). Each message $\mu_{i\leftarrow j}(x_i)$ is updated according to

$$\mu_{i\leftarrow j}^{new}(x_i) = \int dx_j \, \Psi_{ij}(x_i, x_j) \prod_{k\in\partial j\setminus i} \mu_{j\leftarrow k}\left(x_j\right), \tag{2.3}$$

where $\partial i = \{j : j \sim i\}$ denotes the index set of variables connected to x_i in G. At each step the current approximations $q_{ij}(x_{i,j}\,)$ of $p(x_i, x_j)$ can be computed according to

$$q_{ij}\left(x_i, x_j\right) \propto \Psi_{ij}(x_i, x_j) \prod_{k\in\partial j\setminus i} \mu_{j\leftarrow k}\left(x_j\right) \prod_{l\in\partial i\setminus j} \mu_{i\leftarrow l}\left(x_i\right). \tag{2.4}$$

The update steps in (2.11) have to be iterated until convergence. The corresponding $q_{ij}(x_i, x_j)$s yield the final approximation of the $p(x_i, x_j)$s. It is common to use damping, that is, to replace $\mu_{i\leftarrow j}^{new}(x_i)$ by $\mu_{i\leftarrow j}(x_i)^{1-\epsilon}\mu_{i\leftarrow j}^{new}(x_i)^{\epsilon}$ with $\epsilon \in (0, 1]$. In practice this helps to eliminate limit cycles (periodic paths of (2.3)) while it keeps the characteristic of the equilibrium point unchanged. Figure 2.1 illustrates the incoming messages to nodes x_i and x_j. A quite significant difference between the discrete and Gaussian the message passing is the replacement of the sum operator with the integral operator. While finite sums always exist, the integral in (2.3) can become infinite. This problem can be remedied technically by a proper parameterization (see Section 2.4) which keeps the algorithm running, but it can lead to non-normalizable approximate marginals q_{ij}, and thus a (possible) break-down of the algorithm.

Message passing was introduced by Pearl (1988) as a heuristic algorithm (in discrete models), however, Yedidia et al. (2000) showed that it can also be viewed as an algo-

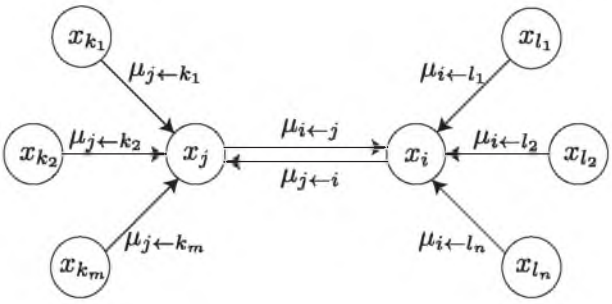

Figure 2.1: An illustration of the incoming messages to nodes x_i and x_j.

rithm for finding the stationary points of the so-called Bethe free energy, an error function measuring the difference between p and a specific family of distributions to be detailed in the next section. It has been shown by Heskes (2003) and later in a different way by Watanabe and Fukumizu (2009) that stable fixed points of the (loopy) message passing algorithm are local minima of the corresponding Bethe free energy. In this chapter we show that this holds for Gaussian models as well.

Our interest in the Gaussian Bethe free energy and the corresponding Gaussian message passing algorithm is motivated by theoretical curiosity: we are interested in the existence of local minima and the convergence properties as well as their implications to more general models and methods like non-Gaussian models and expectation propagation, respectively. For this reason, we will not compare the speed of the method and the accuracy of the approximation with the above mentioned exact linear algebraic methods.

As mentioned in the introduction, the approach we take is similar to that in Welling and Teh (2001), that is, we study the properties of the Gaussian Bethe free energy, parameterized in terms of the moment parameters of the approximate marginals, to find out whether there are local minima to which message passing can converge. In the following we introduce the mean field and the Bethe approximation in Gaussian models. Readers familiar with this subject can continue with Section 2.3.

2.2.1 The Gaussian Bethe free energy

A popular method to approximate marginals is approximating p with a distribution q having a form that makes marginals easy to identify, for example, it factorizes or it has a "tree-like" form. The most common quantity to measure the difference between two probability distributions is the Kullback-Leibler divergence $\mathrm{D}\left[q \,||\, p\right]$. It is often used (e.g. Jaakkola, 2000) to characterize the quality of the approximation and formulate the computation of approximate marginals as the optimization problem

$$q^*(\boldsymbol{x}) = \operatorname*{argmin}_{q \in \mathcal{F}} \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \left[\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right]. \tag{2.5}$$

Here, $\mathcal{F}$ is the set of distributions with the above mentioned form. Since it is not symmetric, the Kullback-Leibler divergence is not a distance, but $\mathrm{D}\,[q\,||\,p] \geq 0$ for any proper q and p, $\mathrm{D}\,[q\,||\,p] = 0$ if and only if $p = q$, and it is convex in both q and p.

A family $\mathcal{F}$ of densities possessing a form that makes marginals easy to identify is the family of distributions that factorize as $q(\boldsymbol{x}) = \prod_k q_k(x_k)$. In other words, in problem (2.5) we approximate p with a distribution that has independent variables. An approximation q of this type is called mean field approximation (e.g. Jaakkola, 2000). Defining $F_{\mathrm{MF}}(\{q_k\}) = \mathrm{D}\,[\prod q_k\,||\,p]$ and writing out in detail the right hand side of (2.5) one gets

$$F_{\mathrm{MF}}(\{q_k\}) = -\int d\boldsymbol{x} \prod_k q_k(x_k) \log p(\boldsymbol{x}) + \sum_k \int dx_k\, q_k(x_k) \log q_k(x_k). \quad (2.6)$$

Using the parameterization $q_k(x_k) = N(x_k|m_k, v_k)$, and the notation $\boldsymbol{m} = (m_1, \ldots, m_n)^T$ and $\boldsymbol{v} = (v_1, \ldots, v_n)^T$, this reduces to

$$F_{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{v}) = -\boldsymbol{h}^T\boldsymbol{m} + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{Q}\boldsymbol{m} + \frac{1}{2}\sum_k Q_{kk}v_k - \frac{1}{2}\sum_k \log(v_k) + C_{\mathrm{MF}}, \quad (2.7)$$

where C_{MF} is an irrelevant constant. Although $\mathrm{D}\,[\prod_k q_k\,||\,p]$ might not be convex in $(q_1, \ldots, q_n)$, one can easily check that F_{MF} is convex in its variables $\boldsymbol{m}$ and $\boldsymbol{v}$ and its minimum is obtained for $\boldsymbol{m} = \boldsymbol{Q}^{-1}\boldsymbol{h}$ and $v_k = 1/Q_{kk}$. Since

$$\left[\boldsymbol{Q}^{-1}\right]_{kk} = \left(Q_{kk} - \boldsymbol{Q}^T_{k,\setminus k}\left[\boldsymbol{Q}_{\setminus k,\setminus k}\right]^{-1}\boldsymbol{Q}_{\setminus k,k}\right)^{-1}$$

one can easily see that the mean field approximation underestimates variances. The mean field approximation computes a solution in which the means are exact, but the variances are computed as if there were no interactions between the variables, namely, as if the matrix $\boldsymbol{Q}$ were diagonal, thus giving poor estimates of the variances.

In order to improve the estimates for variances, one has to choose approximating distributions q that are able to capture dependencies between the variables in p. It can be verified that any distribution in which the dependencies form a tree graph can be written in the form

$$p(\boldsymbol{x}) = \prod_{i\sim j} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \prod_k p(x_k),$$

where i and j run through all the connections or edges (i, j) of the tree and k runs through $\{1, \ldots, n\}$. Although in most cases the undirected graph generated by the non-zero elements in $\boldsymbol{Q}$ is not a tree, based on the "tree intuition" one can construct q from one and two variable marginals as

$$q(\boldsymbol{x}) \propto \prod_{i\sim j} \frac{q_{ij}(x_i, x_j)}{q_i(x_i)q_j(x_j)} \prod_k q_k(x_k) \quad (2.8)$$

and constrain the functions q_{ij} and q_k to be marginally consistent and normalize to 1, that is, $\int dx_j q_{ij}(x_i, x_j) = q_i(x_i)$ for any $i \sim j$ and $\int dx_k q_k(x_k) = 1$ for any k. An approximation of the form (2.8) together with the constraints on q_{ij}s and q_ks is called

Bethe approximation. Let us denote the family of such functions by $\mathcal{F}_B$. By choosing $q_{ij}(x_i, x_j) = q_i(x_i)q_j(x_j)$ one can easily check that $\mathcal{F}_{\mathrm{MF}} \subset \mathcal{F}_B$, thus $\mathcal{F}_B$ is non-empty. Assuming that the approximate marginals are correct and q normalizes to 1 and then substituting (2.8) into (2.5), we get an approximation of the Kullback–Leibler divergence in (2.5) called the Bethe free energy.

Due to the factorization of p, we can write the Bethe free energy as

$$F_B(\{q_{ij}, q_k\}) = -\sum_{i\sim j} \int d\boldsymbol{x}_{i,j}\, q_{ij}(\boldsymbol{x}_{i,j}) \log \Psi_{ij}(\boldsymbol{x}_{i,j}) \tag{2.9}$$
$$+ \sum_{i\sim j} \int d\boldsymbol{x}_{i,j}\, q_{ij}(\boldsymbol{x}_{i,j}) \log \left[\frac{q_{ij}(\boldsymbol{x}_{i,j})}{q_i(x_i)q_j(x_j)}\right] + \sum_k \int dx_k\, q_k(x_k) \log q_k(x_k).$$

One can also define the free energy through the Bethe approximation

$$\int d\boldsymbol{x}\, q(\boldsymbol{x}) \log q(\boldsymbol{x}) \approx \sum_{i\sim j} \int d\boldsymbol{x}_{i,j}\, q(\boldsymbol{x}_{i,j}) \log q(\boldsymbol{x}_{i,j})$$
$$+ \sum_k (1 - n_k) \int dx_k\, q(x_k) \log q(x_k)$$

of the entropy (e.g. Yedidia et al., 2000) and substitute the marginals with functions q_{ij} and q_k that normalize to one and are connected through the marginal consistency constraints $\int dx_j q_{ij}(x_i, x_j) = q_i(x_i)$.

From the stationary conditions of the Lagrangian corresponding to the fractional Bethe free energy (2.9) and the marginal consistency and normalization constraints, one can derive the same iterative algorithm as in (2.3) for the corresponding Lagrange multipliers of the consistency constraints (Yedidia et al., 2000). Similarly, approximate marginals can then be computed according to (2.4). That is, it is easy to show that there is a one-to-one correspondence between the stationary points of the Bethe free energy (2.9) and the fixed points of the message passing algorithm (2.3). Later, in Section 2.4 we will link the stable fixed point of (2.3) to the local minima of (2.9).

2.2.2 Fractional free energies and message passing

As mentioned in the introduction, in case of Gaussian models the message passing algorithm does not always converge, and the reason for this appears to be that the approximate marginals may get indefinite or negative definite covariance matrices. Welling and Teh (2001) pointed out that this can be due to the unboundedness of the Bethe free energy.

Since F_{MF} is convex and bounded and the Bethe free energy might be unbounded, it seems plausible to analyze the fractional Bethe free energy

$$F_{\boldsymbol{\alpha}}(\{q_{ij}, q_k\}) = -\sum_{i\sim j} \int d\boldsymbol{x}_{i,j}\, q_{ij}(\boldsymbol{x}_{i,j}) \log \Psi_{ij}(\boldsymbol{x}_{i,j}) \tag{2.10}$$
$$+ \sum_{i\sim j} \frac{1}{\alpha_{ij}} \int d\boldsymbol{x}_{i,j}\, q_{ij}(\boldsymbol{x}_{i,j}) \log \left[\frac{q_{ij}(\boldsymbol{x}_{i,j})}{q_i(x_i)q_j(x_j)}\right] + \sum_k \int dx_k\, q_k(x_k) \log q_k(x_k).$$

introduced by Wiegerinck and Heskes (2003). Here, $\boldsymbol{\alpha}$ denotes the set of positive reals $\{\alpha_{ij}\}$. They showed that the fractional Bethe free energy "interpolates" between the mean field and the Bethe approximation. That is, for $\alpha_{ij} = 1$ we get the Bethe free energy, while in the case when all α_{ij}s tend to 0, the mutual information between variables x_i and x_j is highly penalized, therefore, (2.10) enforces solutions close to the mean field solution. They also showed that the fractional message passing algorithm derived from (2.10) can be interpreted as Pearl's message passing algorithm with the difference that instead of computing local marginals—like in Pearl's algorithm—one computes local α_{ij}–marginals.[4] The local α_{ij}–marginals correspond to "true" local marginals when $\alpha_{ij} = 1$ and to local mean field approximations when $\alpha_{ij} = 0$. The resulting algorithm is called the fractional message passing algorithm and it reads

$$\mu_{i\leftarrow j}^{new}(x_i)^{\alpha} = \int dx_j \, \Psi_{ij}^{\alpha}(x_i, x_j) \prod_{k\in\partial j\setminus i} \mu_{j\leftarrow k}(x_j) \, \mu_{j\leftarrow i}(x_j)^{1-\alpha}, \tag{2.11}$$

while the approximate marginals are computed according to

$$q_{ij}(x_i, x_j) \propto \Psi_{ij}^{\alpha}(x_i, x_j) \prod_{k\in\partial j\setminus i} \mu_{j\leftarrow k}(x_j) \, \mu_{j\leftarrow i}(x_j)^{1-\alpha} \prod_{l\in\partial i\setminus j} \mu_{i\leftarrow l}(x_i) \, \mu_{i\leftarrow j}(x_i)^{1-\alpha}. \tag{2.12}$$

Power expectation propagation by Minka (2004) is an approximate inference method that uses local approximations with α–divergences. In case of Gaussian models power expectation propagation—with a fully factorized approximating distribution—leads to the same message passing algorithm as the one derived from (2.10) and the appropriate constraints. Starting from the idea of creating an upper bound on the log partition function when p and q are exponential distributions, Wainwright et al. (2003) derived a form of (2.10) where the α_{ij}s are chosen such that this bound is convex in $\{q_{ij}, q_k\}$.

Message passing works well in practice, however, there are other ways to find the local minima of the fractional free energies like the direct minimization w.r.t. some parameterization of the approximate marginals q_{ij} and q_k (Welling and Teh, 2001). The latter method is slower but more likely to converge. In the following we analyze the Bethe free energy when expressed in terms of the moment parameters of the approximate marginals q_{ij}. Later in Section 2.4 we analyze the stability conditions of the fractional message passing algorithm and by expressing these conditions in term of the moment parameters of the approximate marginals we show that stable fixed points of the fractional Gaussian message passing are local minima of the fractional Bethe free energy.

[4] We define the α–marginals of a distribution p as $\operatorname{argmin}_{\{q_k\}} D_{\alpha}\left[p \,\|\, \prod_k q_k\right]$, where D_{α} is the α–divergence $\mathrm{D}_{\alpha}[p\,||\,q] = \left[\int d\boldsymbol{x} p(\boldsymbol{x})^{\alpha} q(\boldsymbol{x})^{1-\alpha} + \alpha \int d\boldsymbol{x} p(\boldsymbol{x}) + (1-\alpha)\int d\boldsymbol{x} q(\boldsymbol{x})\right] / \alpha(1-\alpha)$ (e.g. Minka, 2005).

2.3 Bounds on the Gaussian Bethe free energy

In this section we analyze the parametric form of (2.10). We show that the fractional Gaussian Bethe free energy is a non-increasing function of $\boldsymbol{\alpha}$. By letting all α_{ij} tend to infinity, we obtain a lower bound for the free energies. It turns out that the condition for the lower bound to be bounded from below is the same as the pairwise normalizability condition of Malioutov et al. (2006).

As mentioned in Section 2.2, without loss of generality, we can work with a unit diagonal $\boldsymbol{Q}$. We define $\boldsymbol{R}$ to be a matrix with zeros on its diagonal and $\boldsymbol{Q} = \boldsymbol{I} + \boldsymbol{R}$, where $\boldsymbol{I}$ is the identity matrix. $|\boldsymbol{R}|$ will be the matrix formed by the absolute values of $\boldsymbol{R}$'s elements. We use the moment parameterization $q_{ij}(\boldsymbol{x}_{i,j}) = N(\boldsymbol{x}_{i,j}|\boldsymbol{m}_{ij}, \boldsymbol{V}_{ij})$ and $q_k(x_k) = N(x_k|m_k, v_k)$, where $\boldsymbol{m}_{ij} = (m_{ij}^i, m_{ij}^j)^T$ and $\boldsymbol{V}_{ij} = [v_{ij}^i,\ v_{ij};\ v_{ji},\ v_{ij}^j]$, with $v_{ij} = v_{ji}$. By using $m_i \equiv m_{ij}^i = m_{ik}^i$ and $v_i \equiv v_{ij}^i = v_{ik}^k$ for all $i \sim j$ and $i \sim k$, we embed the marginalization ($\int dx_j q_{ij}(x_i, x_j) = q_i(x_i)$ for all $i \sim j$) and normalization ($\int dx_j q_j(x_j) = 1$) constraints into the parameterization. With a slight abuse of notation the matrix formed by diagonal elements v_k and off-diagonal elements v_{ij} is denoted by $\boldsymbol{V}$ (we can take $v_{ij} = 0$ for all $i \nsim j$), the vector of means by $\boldsymbol{m} = (m_1, \ldots, m_n)^T$ and the vector of variances by $\boldsymbol{v} = (v_1, \ldots, v_n)^T$. Substituting q_{ij} and q_k into (2.10) one gets

$$\begin{aligned} F_{\boldsymbol{\alpha}}(\boldsymbol{m}, \boldsymbol{V}) = &- \boldsymbol{h}^T\boldsymbol{m} + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{Q}\boldsymbol{m} + \frac{1}{2}\mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{V}) \\ &- \frac{1}{2}\sum_{i\sim j}\frac{1}{\alpha_{ij}}\log\left(1 - \frac{v_{ij}^2}{v_i v_j}\right) - \frac{1}{2}\sum_k \log(v_k) + C, \end{aligned} \tag{2.13}$$

where C is an irrelevant constant. Note that the variables $\boldsymbol{m}$ and $\boldsymbol{V}$ are independent, hence the minimizations of $F_{\boldsymbol{\alpha}}(\boldsymbol{m}, \boldsymbol{V})$ with regard to $\boldsymbol{m}$ and $\boldsymbol{V}$ can be carried out independently.

Property 1. *$F_{\boldsymbol{\alpha}}(\boldsymbol{m}, \boldsymbol{V})$ is convex and bounded in $(\boldsymbol{m}, \{v_{ij}\}_{i\neq j})$ and at any stationary point we have*

$$\begin{aligned} \boldsymbol{m}^* &= \boldsymbol{Q}^{-1}\boldsymbol{h} \\ v_{ij}^* &= -\mathrm{sign}(R_{ij})\frac{\sqrt{1 + (2\alpha_{ij}R_{ij})^2 v_i v_j} - 1}{2\alpha_{ij}|R_{ij}|}. \end{aligned} \tag{2.14}$$

Proof: By definition, $\boldsymbol{Q}$ is positive definite, therefore, the quadratic term in $\boldsymbol{m}$ is convex and bounded. The variables $\boldsymbol{m}$ and $\boldsymbol{V}$ are independent and the minimum with regard to $\boldsymbol{m}$ is achieved at $\boldsymbol{m}^* = \boldsymbol{Q}^{-1}\boldsymbol{h}$. One can check that the second order derivative of $F_{\boldsymbol{\alpha}}(\boldsymbol{m}, \boldsymbol{V})$ with regard to v_{ij} is non-negative and the first order derivative has only one solution when $-v_i v_j \leq v_{ij}^2 \leq v_i v_j$. Since the variables v_{ij} are independent, one can conclude that $F_{\boldsymbol{\alpha}}(\boldsymbol{m}, \boldsymbol{V})$ is convex in v_{ij}. From the independence of $\boldsymbol{m}$ and $\boldsymbol{V}$, it follows that $F_{\boldsymbol{\alpha}}$ is convex in $(\boldsymbol{m}, \{v_{ij}\}_{i\neq j})$. ■

Since the $\boldsymbol{V}_{ij}$s are constrained to be covariance matrices, we have $v_i v_j > v_{ij}^2$, thus the first logarithmic term in (2.13) is negative. As a consequence,

$$F_{\boldsymbol{\alpha}_1}(\boldsymbol{m}, \boldsymbol{V}) \geq F_{\boldsymbol{\alpha}_2}(\boldsymbol{m}, \boldsymbol{V}) \quad \text{for any} \quad 0 < \boldsymbol{\alpha}_1 \leq \boldsymbol{\alpha}_2,$$

where $\boldsymbol{\alpha}_1 \leq \boldsymbol{\alpha}_2$ is taken element by element. This observation leads to the following property.

Property 2. *With $\alpha_{ij} = \alpha$, F_α is a non-increasing function of α.*

Using Property 1 and substituting v_{ij}^* into $F_{\boldsymbol{\alpha}}$ we define the constrained function

$$\begin{aligned} F_{\boldsymbol{\alpha}}^c(\boldsymbol{m}, \boldsymbol{v}) = & -\boldsymbol{h}^T\boldsymbol{m} + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{Q}\boldsymbol{m} + \frac{1}{2}\sum_k v_k \\ & - \frac{1}{2}\sum_{i\sim j}\frac{1}{\alpha_{ij}}\left(\sqrt{1+(2\alpha_{ij}R_{ij})^2 v_i v_j} - 1\right) \\ & - \frac{1}{2}\sum_{n(i,j)}\frac{1}{\alpha_{ij}}\log\left(2\frac{\sqrt{1+(2\alpha_{ij}R_{ij})^2 v_i v_j} - 1}{(2\alpha_{ij}R_{ij})^2 v_i v_j}\right) \\ & - \frac{1}{2}\sum_k \log(v_k) + C^c, \end{aligned} \tag{2.15}$$

where C^c is an irrelevant constant. From Property 2, it follows that when choosing $\alpha_{ij} = \alpha$, the function in (2.15) is a non-increasing function of α. It then makes sense to take $\alpha \to \infty$ and verify whether we can get a lower bound for (2.15).

Lemma 1. *For any $\boldsymbol{v} > 0$, $0 \leq \alpha_1 \leq 1$ and $\alpha_2 \geq 1$ the following inequalities hold.*

$$\begin{aligned} F_{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{v}) &\geq F_{\alpha_1}^c(\boldsymbol{m}, \boldsymbol{v}) \geq F_B\left(\boldsymbol{m}, \{v_{ij}^*\}, \boldsymbol{v}\right) \\ F_B\left(\boldsymbol{m}, \{v_{ij}^*\}, \boldsymbol{v}\right) &\geq F_{\alpha_2}^c(\boldsymbol{m}, \boldsymbol{v}) \dots \\ \dots &\geq F_{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{v}) - \frac{1}{2}\sqrt{\boldsymbol{v}}^T|\boldsymbol{R}|\sqrt{\boldsymbol{v}} \end{aligned}$$

Moreover, they are tight, that is,

$$\lim_{\alpha\to 0} F_\alpha\left(\boldsymbol{m}, \{v_{ij}^*(\alpha)\}, \boldsymbol{v}\right) = F_{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{v})$$

and

$$\lim_{\alpha\to\infty} F_\alpha\left(\boldsymbol{m}, \{v_{ij}^*(\alpha)\}, \boldsymbol{v}\right) = F_{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{v}) - \frac{1}{2}\sqrt{\boldsymbol{v}}^T|\boldsymbol{R}|\sqrt{\boldsymbol{v}}.$$

Proof: Since the Bethe free energy is the specific case of the fractional Bethe free energy for $\alpha = 1$, the inequalities on $F_B(\boldsymbol{m}, \{v_{ij}^*(\alpha)\}, \boldsymbol{v})$ follow from Property 2. Now, we show that the upper and lower bounds are tight. The function $(1+x^2)^{1/2} - 1$ behaves as $\frac{1}{2}x^2$ in the neighborhood of 0, therefore,

$$\lim_{\alpha\to 0} v_{ij}^*(\alpha) = 0 \qquad \text{and} \qquad \lim_{\alpha\to 0}\frac{\log\left(1 - \frac{{v_{ij}^*}^2(\alpha)}{v_i v_j}\right)}{\alpha} = -\frac{1}{v_i v_j}\lim_{\alpha\to 0}\frac{{v_{ij}^*}^2(\alpha)}{\alpha} = 0,$$

showing that $F_{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{v})$ is a tight upper bound.

As α tends to infinity, we have

$$\lim_{\alpha\to\infty} \frac{\sqrt{1+(2\alpha R_{ij})^2 v_i v_j}-1}{2\alpha} = |R_{ij}|\sqrt{v_i}\sqrt{v_j}$$

and

$$\lim_{\alpha\to\infty} \frac{1}{\alpha}\log\left(\frac{\sqrt{1+(2\alpha R_{ij})^2 v_i v_j}-1}{(2\alpha R_{ij})^2 v_i v_j}\right) = 0,$$

yielding a tight lower bound

$$\lim_{\alpha\to\infty} F_\alpha\left(\boldsymbol{m}, \{v_{ij}^*(\alpha)\}, \boldsymbol{v}\right) = F_{\mathrm{MF}}\left(\boldsymbol{m}, \boldsymbol{v}\right) - \frac{1}{2}\sqrt{\boldsymbol{v}}^T|\boldsymbol{R}|\sqrt{\boldsymbol{v}}. \qquad \blacksquare$$

Let $\lambda_{max}(|\boldsymbol{R}|)$ be the largest eigenvalue of $|\boldsymbol{R}|$. Analyzing the boundedness of the lower bound, we arrive at the following theorem.

Theorem 1. *For the fractional Bethe free energy in (2.13) corresponding to a connected Gaussian model, the following statements hold*

(1) if $\lambda_{max}(|\boldsymbol{R}|) < 1$, *then* $F_{\boldsymbol{\alpha}}$ *is bounded from below for all* $\boldsymbol{\alpha} > 0$,

(2) if $\lambda_{max}(|\boldsymbol{R}|) > 1$, *then* $F_{\boldsymbol{\alpha}}$ *is unbounded from below for all* $\boldsymbol{\alpha} > 0$,

(3) if $\lambda_{max}(|\boldsymbol{R}|) = 1$, *then* $F_{\boldsymbol{\alpha}}$ *is bounded from below if and only if* $\sum_i \sum_{i\sim j} \alpha_{ij}^{-1} \geq 2n$.

Proof: Since in $F_{\boldsymbol{\alpha}}$ there is no interaction between the parameters $\boldsymbol{m}$ and $\boldsymbol{V}$ and the term depending on $\boldsymbol{m}$ is bounded from below due to the positive definiteness of $\boldsymbol{Q}$, we can simply neglect this term when analyzing the boundedness of $F_{\boldsymbol{\alpha}}$. Let us write out in detail the lower bound of the fractional Bethe free energies in the form

$$F_{\mathrm{MF}}\left(\boldsymbol{m}, \boldsymbol{v}\right) - \frac{1}{2}\sqrt{\boldsymbol{v}}^T|\boldsymbol{R}|\sqrt{\boldsymbol{v}} = \tag{2.16}$$
$$\frac{1}{2}\boldsymbol{m}^T\boldsymbol{Q}^{-1}\boldsymbol{m} - \boldsymbol{h}^T\boldsymbol{m} + \frac{1}{2}\sqrt{\boldsymbol{v}}^T\left(\boldsymbol{I}-|\boldsymbol{R}|\right)\sqrt{\boldsymbol{v}} - \frac{1}{2}\mathbf{1}^T\log(\boldsymbol{v}) + \text{const.}$$

Statement (1): The condition $\lambda_{max}(|\boldsymbol{R}|) < 1$ implies that $\boldsymbol{I} - |\boldsymbol{R}|$ is positive definite. Now, $\log(x) \leq x - 1$, thus $\frac{1}{2}\sqrt{\boldsymbol{v}}^T(\boldsymbol{I}-|\boldsymbol{R}|)\sqrt{\boldsymbol{v}} - \mathbf{1}^T\log(\sqrt{\boldsymbol{v}}) \geq \frac{1}{2}\sqrt{\boldsymbol{v}}^T(\boldsymbol{I}-|\boldsymbol{R}|)\sqrt{\boldsymbol{v}} - \mathbf{1}^T\sqrt{\boldsymbol{v}} + n$. The latter is bounded from below and so it follows that (2.16) is bounded from below as well. According to Lemma 1, the boundedness of (2.16) implies that all fractional Bethe free energies are bounded from below.

Statement (2): Since we assumed that the Gaussian network is connected and undirected, it follows that $|\boldsymbol{R}|$ is irreducible (e.g. Horn and Johnson, 2005). According to the Perron-Frobenius theory of non-negative matrices (e.g. Horn and Johnson, 2005), the non-negative and irreducible matrix $|\boldsymbol{R}|$ has a simple maximal eigenvalue $\lambda_{max}(|\boldsymbol{R}|)$ and all elements of the eigenvector $\boldsymbol{u}_{max}$ corresponding to it are positive. Let us take the fractional Bethe free energy and analyze its behavior when $\sqrt{\boldsymbol{v}} = t\boldsymbol{u}_{max}$ and $t \to \infty$. For large values of t we have $(1 + (2\alpha_{ij}R_{ij})^2(u_{max}^i u_{max}^j)^2 t^4)^{1/2} \simeq 2\alpha_{ij}|R_{ij}|u_{max}^i u_{max}^j t^2$,

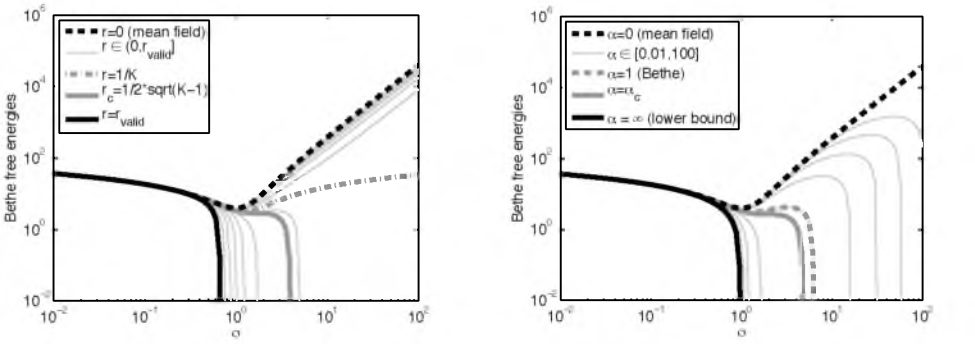

Figure 2.2: Visualizing critical parameters for a symmetric K-regular Gaussian model with off-diagonal elements $R_{ij} = r$. Plots in the left panel correspond to the constrained fractional Bethe free energies F_α^c for $\sqrt{\boldsymbol{v}} = \sigma \mathbf{1}$ for an 8 node 4–regular Gaussian model with r=0.27 ($Kr > 1$) and varying α. Plots in the right panel correspond to the constrained Bethe free energies F_1^c for $\sqrt{\boldsymbol{v}} = \sigma \mathbf{1}$ in an 8 node 4–regular Gaussian model with varying r. Here, r_{valid} is the supremum of rs for which the model is valid, that is, $\boldsymbol{Q}$ is positive definite.

therefore, the sum of the second and third term in (2.15) simplifies to $(1 - \lambda_{max}(|\boldsymbol{R}|))t^2$ and this term dominates over the logarithmic ones as $t \to \infty$. As a result, the limit is independent of the choice of α_{ij} and it tends to $-\infty$ whenever $\lambda_{max}(|\boldsymbol{R}|) > 1$.
Statement (3): If $\lambda_{max}(|\boldsymbol{R}|) = 1$, then the only direction in which the quadratic term will not dominate is $\sqrt{\boldsymbol{v}} = t\boldsymbol{u}_{max}$. Therefore, we have to analyze the behavior of the logarithmic terms in (2.15) when $t \to \infty$. For large ts these behave as $(\sum_{i\sim j} \alpha_{ij}^{-1} - 2n)\log(t)$. For this reason, the boundedness of $F_{\boldsymbol{\alpha}}^c$—and thus of $F_{\boldsymbol{\alpha}}$—depends on the condition in statement (3). ■

It was shown by Malioutov et al. (2006) that the condition $\lambda_{max}(|\boldsymbol{R}|) < 1$ is an equivalent condition to pairwise normalizability. Therefore, pairwise normalizability is not only a sufficient condition for the message passing algorithm to converge, but it is also a necessary condition for the fractional Gaussian Bethe free energies to be bounded. Using Lemma 1, we can show that for a suitable chosen $\epsilon > 0$ there always exists an α_ϵ such that the constrained fractional free energy F_α^c possesses a local minimum for any $0 < \alpha < \alpha_\epsilon$ (Property A2 in Section A.1 of the Appendix).

Example In the case of models with an adjacency matrix (non-zero entries of $\boldsymbol{R}$) corresponding to a K–regular graph[5] and equal interaction weights $R_{ij} = r$, the maximal eigenvalue of $|\boldsymbol{R}|$ is $\lambda_{max}(|\boldsymbol{R}|) = Kr$ and the eigenvector corresponding to this eigenvalue is $\mathbf{1}$. (We define $\mathbf{1}$ as the vector that has all its elements equal to 1.) Verifying the stationary point conditions, it turns out that for some choice of r and α there exists a local minimum which is symmetrical, that is, it lies in the direction $\mathbf{1}$. One can show that when the model is not pairwise normalizable ($Kr > 1$), the critical r below which the fractional Bethe free energy possesses this local minimum is $r_c(K, \alpha) = 1/2\sqrt{\alpha(K - \alpha)}$ and for any valid r the critical α below which the fractional Bethe free energies possesses this local minimum is $\alpha_c(K, r) = \frac{1}{2}K(1 - \sqrt{1 - 1/(Kr)^2})$. These results are illustrated in

[5] A K–regular graph is a graph where all nodes are connected to K other nodes.

Figure 2.2. (Note that for 2–regular graphs, all valid models are pairwise normalizable and possess a unique global minimum.) ■

For K–regular graphs, the convexity of the fractional Bethe free energy in terms of $\{q_{ij}, q_k\}$ requires $\alpha \geq K$, a much stronger condition than $\alpha \geq \alpha_c(K, r)$. Thus, if we choose α sufficiently large such that the Bethe free energy is guaranteed to have a unique global minimum, this minimum is unbounded.

This example disproves the conjecture in Welling and Teh (2001), that is, even when the Bethe free energy is not bounded from below, it can possess a finite local minimum to which the message passing and the minimization algorithms can converge.

2.4 Message passing in Gaussian models

In this section, we turn our attention towards the properties of the message passing algorithm in Gaussian models. Following a similar line of argument as Watanabe and Fukumizu (2009) we show that stable fixed—or equilibrium—points of the message passing algorithm correspond to local minima of the Bethe free energy. The line of reasoning is similar to that of Watanabe and Fukumizu (2009), however, for the Gaussian case, we have to come up with a specific parameterization to be able to follow the same arguments. This parameterization will be the moment parameterization introduced in the previous chapters and also in Cseke and Heskes (2008). The way we proceed is the following: (1) we make a linear expansion of message passing iteration at a fixed point, (2) we express the linear expansion in terms of moment parameters corresponding to the fixed point and finally (3) we connect the properties of the latter with the properties of the Hessian of the Bethe free energy by using the matrix determinant lemma (Watanabe and Fukumizu, 2009).

The form of the equation (2.11) implies that the messages $\mu_{i \leftarrow j}(x_i)$ are univariate Gaussian functions, thus we can express them in terms of two scalar (canonical) parameters η_{ij} and λ_{ij} such that $\log \mu_{i \leftarrow j}(x_i) = -\lambda_{ij} x_i^2/2 + \eta_{ij} x_i$. When expressed in terms of η_{ij} and λ_{ij}, the damped message passing algorithm (2.11) translates to

$$\eta_{ij}^{new} = (1-\epsilon)\eta_{ij} + \frac{\epsilon}{\alpha}\left[\alpha\gamma_{ij}^i h_i - \alpha R_{ij}\frac{\alpha\gamma_{ij}^j h_j + \sum_{k\in\partial j\setminus i}\eta_{jk} + (1-\alpha)\eta_{ji}}{\alpha\gamma_{ij}^j + \sum_{k\in\partial j\setminus i}\lambda_{jk} + (1-\alpha)\lambda_{ji}}\right] \quad (2.17)$$

$$\lambda_{ij}^{new} = (1-\epsilon)\lambda_{ij} + \frac{\epsilon}{\alpha}\left[\alpha\gamma_{ij}^i - \alpha^2 R_{ij}^2\left(\alpha\gamma_{ij}^j + \sum_{k\in\partial j\setminus i}\lambda_{jk} + (1-\alpha)\lambda_{ji}\right)^{-1}\right] \quad (2.18)$$

where $\gamma_{ij}^i, \gamma_{ij}^j, h_i$ and R_{ij} are parameters of Ψ_{ij} as in Section 2.2.1, with $R_{ij} = Q_{ij}$ and the assumption that $Q_{ii} = 1$. The approximate marginals q_{ij} in (2.12) might not be normalizable, but the message passing iteration in (2.17) and (2.18) stays well defined unless there is a zero in the denominator on the rhs. This rarely happens in practice. However, it is more common that message passing converges while there are some intermediate steps at which the approximate marginals q_{ij} are not normalizable. This can often be remedied by choosing an appropriate damping parameter ϵ. The complexity of one message update scales roughly with the number of (directed) edges times the average number of edges a node has, that is, $nnzeros(\boldsymbol{Q})^2/n$.

The iteration (2.18) for the λ_{ij}s is independent of η_{ij}s and the iteration (2.17) for the η_{ij}s is linear in η_{ij}. It is interesting to see that when $\boldsymbol{h} = 0$ neither the constrained Bethe free energy (2.15) nor the message passing algorithm (2.18) depend on the sign of R_{ij}. These are only relevant to compute the means—when $\boldsymbol{h} \neq 0$—and the correlations in (2.14). As a result, the marginal variances computed by either minimizing the Bethe free energy or by running the message passing algorithm can only depend on $|\boldsymbol{R}|$, similarly to the constrained fractional free energy F_α^c.

2.4.1 Stability of Gaussian message passing

In the following we analyze the stability of the message passing iteration at its fixed points, that is, at the stationary points of the Lagrangian corresponding to the constrained minimization of the Gaussian Bethe free energy. We reiterate that we use $G = (V, E)$ to denote the graph corresponding to $\boldsymbol{Q}$, namely, $V = \{1, \ldots, n\}$ and $E = \{(i, j) : Q_{ij} \neq 0\}$. The vector $\boldsymbol{\lambda} \in \mathbb{R}^{|E|}$, corresponding to a set of messages $\{\lambda_{ij}\}_{ij}$, is composed by the concatenation of λ_{ij}s such that ij is followed by ji and the (ij, ji) blocks follow a lexicographic order w.r.t. ij and $i < j$. The vector $\boldsymbol{\eta}$ consists of the variables η_{ij} and follows a similar structure as $\boldsymbol{\lambda}$. We define $\hat{\boldsymbol{r}}, \hat{\boldsymbol{h}}, \hat{\boldsymbol{\gamma}} \in \mathbb{R}^{|E|}$ as $\hat{r}_{ij} = \hat{r}_{ji} = R_{ij}$, $\hat{h}_{ij} = h_j$ and $\hat{\gamma}_{ij} = \gamma_{ij}^j$. We also define the $|E| \times |E|$ matrix

$$\mathcal{M}_{ij,kl}(\alpha) \equiv \begin{cases} 1 & \text{if } j = k \\ 1 - \alpha & \text{if } kl = ji \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

which encodes the weighted edge adjacency corresponding to G and α. The number of non-zero elements in $\boldsymbol{\mathcal{M}}(\alpha)$, scales roughly with $nnzeros(\boldsymbol{Q})^2/n$. As a consequence, the complexity of a parallel update given by Equations (2.17) and (2.18), also scales as roughly with $nnzeros(\boldsymbol{Q})^2/n$.

With this notation, the local linearization of the update equations (2.17) and (2.18) can be written as

$$\frac{\partial(\boldsymbol{\eta}^{new}, \boldsymbol{\lambda}^{new})}{\partial(\boldsymbol{\eta}, \boldsymbol{\lambda})}(\boldsymbol{\eta}, \boldsymbol{\lambda}) = (1 - \epsilon)\boldsymbol{I} \tag{2.20}$$
$$+ \frac{\epsilon}{\alpha} \begin{bmatrix} -\text{diag}\left(\alpha\hat{\boldsymbol{r}}\frac{1}{\alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda}}\right)\boldsymbol{\mathcal{M}}(\alpha) & \text{diag}\left(\alpha\hat{\boldsymbol{r}}\frac{\alpha\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{h}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\eta}}{(\alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda})^2}\right)\boldsymbol{\mathcal{M}}(\alpha) \\ \boldsymbol{0} & \text{diag}\left(\alpha^2\hat{\boldsymbol{r}}^2\frac{1}{(\alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda})^2}\right)\boldsymbol{\mathcal{M}}(\alpha) \end{bmatrix},$$

where all operations on vectors are element by element. The stability of a fixed point $(\boldsymbol{\eta}^*, \boldsymbol{\lambda}^*)$ depends on the union of the spectra of

$$\boldsymbol{J}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*, \boldsymbol{\lambda}^*) \equiv -\alpha^{-1}\text{diag}\left(\alpha\hat{\boldsymbol{r}}(\alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda}^*)^{-1}\right)\boldsymbol{\mathcal{M}}(\alpha)$$

and

$$\boldsymbol{J}_{\boldsymbol{\lambda}}(\boldsymbol{\eta}^*, \boldsymbol{\lambda}^*) \equiv \alpha^{-1}\text{diag}\left(\alpha^2\hat{\boldsymbol{r}}^2(\alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda}^*)^{-2}\right)\boldsymbol{\mathcal{M}}(\alpha).$$

It is important to point out that the stability properties depend only on $\boldsymbol{\lambda}^*$ and $\boldsymbol{R}$ and are

independent of $\boldsymbol{\eta}^*$ and $\boldsymbol{h}$.

Our goal is to connect the stability properties of the message passing algorithm to the properties of the Bethe free energy. Therefore, we express the stability properties in terms of the moment parameters of approximate marginals. For any $\boldsymbol{\lambda}$ that leads to normalizable approximate marginals $q_{ij}(x_i, x_j)$, we can use (2.12) to identify the local covariance parameters $\boldsymbol{V}_{ij}$ defined in Section 2.3, but now without enforcing the marginal matching constraints $v_{ij}^i = v_{ik}^i$. The correspondence is given by

$$\begin{aligned}\begin{bmatrix} v_{ij}^i & v_{ij} \\ v_{ij} & v_{ij}^j \end{bmatrix}^{-1} &= \frac{1}{v_{ij}^i v_{ij}^j - v_{ij}^2} \begin{bmatrix} v_{ij}^j & -v_{ij} \\ -v_{ij} & v_{ij}^i \end{bmatrix} \\ &= \begin{bmatrix} \alpha\gamma_{ij}^i + \sum\limits_{l\in\partial i\setminus j} \lambda_{il} + (1-\alpha)\lambda_{ij} & \alpha R_{ij} \\ \alpha R_{ij} & \alpha\gamma_{ij}^j + \sum\limits_{k\in\partial j\setminus i} \lambda_{jk} + (1-\alpha)\lambda_{ji} \end{bmatrix}.\end{aligned} \tag{2.21}$$

The approximate local covariances v_{ij} are fully determined by v_{ij}^i, v_{ij}^j and r_{ij} and have the form as in (2.14). This leaves us with $|E|$ moment parameters to be computed by the message passing algorithm. Let $\hat{\boldsymbol{v}} \in \mathbb{R}^{|E|}$ be defined as $\hat{v}_{ij} = v_{ij}^i, \hat{v}_{ji} = v_{ij}^j$ and $y_{ij}(\hat{\boldsymbol{v}}) = v_{ij}/(v_{ij}^i v_{ij}^j - {v_{ij}}^2)$, where v_{ij} is computed according to (2.14). It can be checked that the mapping between $\boldsymbol{y}$ and $\hat{\boldsymbol{v}}$ is continuous and bijective. This implies that the canonical to moment parameter transformation in (2.21) can be written as $\boldsymbol{y}(\hat{\boldsymbol{v}}) = \alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda}$. Since $\boldsymbol{\mathcal{M}}(\alpha)$ is singular only when $\alpha = K$ and the graph G is K-regular—see Property A1 in Section A.1 of the Appendix for details—for the rest of the cases, there is a continuous, bijective mapping between the moment parameters $\hat{\boldsymbol{v}}$ and the canonical parameters $\boldsymbol{\lambda}$ that lead to normalizable approximate marginals.

At any fixed point $(\boldsymbol{\eta}^*, \boldsymbol{\lambda}^*)$ we have moment matching, that is, $v_{ij}^i = v_{ik}^i \equiv v_i^*$ for any $k, j \in \partial i$, therefore we can express the stability properties in terms of moment parameters $\boldsymbol{v}^* = (v_i^*, \ldots, v_n^*)$. Using (2.21) and defining the diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{|E|\times|E|}$ with the diagonal elements $D_{ij,ij} = \sqrt{v_i^*}$, we get

$$\boldsymbol{D}\boldsymbol{J}_{\boldsymbol{\eta}}(\boldsymbol{\lambda}^*(\boldsymbol{v}^*))\boldsymbol{D}^{-1} = -\alpha^{-1}\mathrm{diag}\left(\frac{v_{ij}(\alpha, v_i^*, v_j^*)}{\sqrt{v_i^* v_j^*}}\right)\boldsymbol{\mathcal{M}}(\alpha) \tag{2.22}$$

and

$$\boldsymbol{D}^2\boldsymbol{J}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*(\boldsymbol{v}^*))\boldsymbol{D}^{-2} = \alpha^{-1}\mathrm{diag}\left(\frac{v_{ij}(\alpha, v_i^*, v_j^*)^2}{v_i^* v_j^*}\right)\boldsymbol{\mathcal{M}}(\alpha). \tag{2.23}$$

Let $\sigma(\boldsymbol{A})$ denote the spectrum of the matrix $\boldsymbol{A}$. Since we have $\sigma\left(\boldsymbol{D}\boldsymbol{J}_{\boldsymbol{\eta}}\boldsymbol{D}^{-1}\right) = \sigma\left(\boldsymbol{J}_{\boldsymbol{\eta}}\right)$ and $\sigma\left(\boldsymbol{D}^2\boldsymbol{J}_{\boldsymbol{\lambda}}\boldsymbol{D}^{-2}\right) = \sigma\left(\boldsymbol{J}_{\boldsymbol{\lambda}}\right)$, it is sufficient to analyze the spectral properties of the right hand sides in equations (2.22) and (2.23).

The message passing algorithm is asymptotically stable at $\boldsymbol{\lambda}^*(\boldsymbol{v}^*)$ if and only if

$$\max\left\{\rho\left(\boldsymbol{J}_{\boldsymbol{\eta}}(\boldsymbol{\lambda}^*(\boldsymbol{v}^*))\right), \rho\left(\boldsymbol{J}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*(\boldsymbol{v}^*))\right)\right\} < 1, \tag{2.24}$$

where $\rho(\cdot)$ denotes the spectral radius. It is interesting to see that although the functional forms of the free energies and the message passing algorithms are different in the Gaussian and discrete case (Watanabe and Fukumizu, 2009), the stability conditions have similar forms. This will allow us to use some of the results in Watanabe and Fukumizu (2009). In the next section, we show the implications of this condition for the properties of the Hessian of the free energy.

2.4.2 Stable fixed points and local minima

The Hessian $\boldsymbol{H}[F_\alpha]$ of the Bethe free energy (2.13) depends only on the moment parameters v_i, v_j and v_{ij}. Note that now, the v_{ij}s are unconstrained parameters. It is an $(|E|/2+2n)\times(|E|/2+2n)$ matrix and it has the form

$$\boldsymbol{H}[F_\alpha](\boldsymbol{V}) = \begin{bmatrix} \boldsymbol{Q} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \operatorname{diag}\left(\frac{\partial^2 F_\alpha}{\partial^2 v_{ij}}\right) & \left[\frac{\partial^2 F_\alpha}{\partial v_{ij}\partial v_i}\right]_{ij,i} \\ \mathbf{0} & \left[\frac{\partial^2 F_\alpha}{\partial v_{ij}\partial v_i}\right]^T_{ij,i} & \left[\frac{\partial^2 F_\alpha}{\partial v_i \partial v_j}\right]_{i,j} \end{bmatrix},$$

where we use $\boldsymbol{V}$ to denote the collection of parameters $v_i,\ i=1,\ldots,n$ and $v_{ij},\ i\sim j$. Since the block corresponding to the partial differentials w.r.t. v_{ij} is diagonal with positive elements, the Hessian is positive definite at $\boldsymbol{V}$ if the Schur complement corresponding to the partial differentials w.r.t. v_is is positive definite at $\boldsymbol{V}$. The latter is given by

$$\begin{aligned} H^v_{ii}[F_\alpha](\boldsymbol{V}) &= \frac{\partial^2 F_\alpha}{\partial v_i \partial v_i} - \sum_{i\sim j}\left[\frac{\partial^2 F_\alpha}{\partial v_{ij}\partial v_i}\right]^2\left[\frac{\partial F_\alpha}{\partial v_{ij}}\right]^{-1} \\ &= \frac{1}{2}\frac{1}{v_i^2}\left(1+\frac{1}{\alpha}\sum_{i\sim j}\frac{c_{ij}^4}{1-c_{ij}^4}\right) \\ H^v_{ij}[F_\alpha](\boldsymbol{V}) &= \frac{\partial^2 F_\alpha}{\partial v_i \partial v_j} - \frac{\partial^2 F_\alpha}{\partial v_{ij}\partial v_i}\frac{\partial^2 F_\alpha}{\partial v_{ij}\partial v_j}\left[\frac{\partial^2 F_\alpha}{\partial^2 v_{ij}}\right]^{-1} \\ &= -\frac{1}{2}\frac{1}{v_i v_j}\frac{1}{\alpha}\frac{c_{ij}^2}{1-c_{ij}^4} \end{aligned}$$

where we use the notation $c_{ij} = v_{ij}/\sqrt{v_i v_j}$.

Now, we would like to connect the condition in (2.24) to the positive definiteness of the matrix $\boldsymbol{H}^v[F_\alpha](\boldsymbol{V})$. We follow the same line of argument as Watanabe and Fukumizu (2009) and show that stable fixed points $\boldsymbol{\lambda}^*(\boldsymbol{v}^*)$ of the Gaussian message passing algorithm, satisfying (2.24), correspond to local minima of the Gaussian free energy F_α at $\boldsymbol{v}^*$ and $v_{ij}(\alpha, v_i^*, v_j^*)$.

According to Watanabe and Fukumizu (2009), for any arbitrary vector $\boldsymbol{w}\in\mathbb{R}^{|E|}$ one has

$$\det\left(\boldsymbol{I}_{|E|} - \alpha^{-1}\operatorname{diag}(\boldsymbol{w})\,\boldsymbol{\mathcal{M}}(\alpha)\right) = \det\left(\boldsymbol{I}_n + \alpha^{-1}\boldsymbol{A}(\boldsymbol{w})\right)\prod_{ij}(1-w_{ij}w_{ji}), \quad (2.25)$$

where

$$A_{ii}(\boldsymbol{w}) = \sum_{i\sim j} \frac{w_{ij}w_{ji}}{1 - w_{ij}w_{ji}} \qquad \text{and} \qquad A_{ij}(\boldsymbol{w}) = -\frac{w_{ij}}{1 - w_{ij}w_{ji}}. \tag{2.26}$$

The proof is an application of the matrix determinant lemma and a reproduction of it can be found in Section A.1 of the Appendix. Equation (2.25) expresses the determinant of an $|E| \times |E|$ matrix as the determinant of an $n \times n$ matrix.

Let $\boldsymbol{c} \in \mathbb{R}^{|E|}$ with $c_{ij}(\boldsymbol{V}) = v_{ij}/\sqrt{v_i v_j}$. By substituting $\boldsymbol{w} = \boldsymbol{c}(\boldsymbol{V})^2$ in (2.26), we find that

$$\det\left(\boldsymbol{I} - \alpha^{-1}\mathrm{diag}\left(\boldsymbol{c}(\boldsymbol{V})^2\right)\boldsymbol{\mathcal{M}}(\alpha)\right) = f(\boldsymbol{V})\det\left(\boldsymbol{H}[F_\alpha](\boldsymbol{V})\right), \tag{2.27}$$

where $f(\boldsymbol{V})$ is a positive function defined as

$$f(\boldsymbol{V}) = 2^n \alpha^{|E|} |\boldsymbol{Q}|^{-1} \prod_k v_k^2 \prod_{i\sim j} \frac{\left(v_i v_j - v_{ij}^2\right)^2}{v_i v_j + v_{ij}^2} \left(1 - \frac{v_{ij}^2}{v_i v_j}\right).$$

for all $\boldsymbol{V}$ corresponding to normalizable approximate marginals. Now, adapting the theorem of Watanabe and Fukumizu (2009) we have the following theorem.

Theorem (Watanabe and Fukumizu, 2009) *If* $\sigma\left(\alpha^{-1}\mathrm{diag}\left(\boldsymbol{c}(\boldsymbol{V})^2\right)\boldsymbol{\mathcal{M}}(\alpha)\right) \subseteq \mathbb{C} \setminus \mathbb{R}_{\geq 1}$ *then the Hessian of the (Gaussian Bethe) free energy* $\boldsymbol{H}[F_\alpha]$ *is positive definite at* $\boldsymbol{V}$.

Proof: The assumption $\sigma\left(\alpha^{-1}\mathrm{diag}\left(\boldsymbol{c}(\boldsymbol{V})^2\right)\boldsymbol{\mathcal{M}}(\alpha)\right) \subset \mathbb{C} \setminus \mathbb{R}_{\geq 1}$ implies that we have $\det\left(\boldsymbol{I} - \alpha^{-1}\mathrm{diag}(\boldsymbol{c}(\boldsymbol{V})^2)\,\boldsymbol{\mathcal{M}}(\alpha)\right) > 0$. By choosing $V_{ij}(t) = tv_{ij}$ with $t \in [0,1]$, we find that $\boldsymbol{c}(\boldsymbol{V}(t))^2 = t^2\boldsymbol{c}(\boldsymbol{V})^2$, therefore, $\det\left(\boldsymbol{I} - \alpha^{-1}\mathrm{diag}(\boldsymbol{c}(\boldsymbol{V}(t))^2)\boldsymbol{\mathcal{M}}(\alpha)\right) > 0$ for any $t \in [0,1]$. This implies that $\det\left(\boldsymbol{H}[F_\alpha](\boldsymbol{V}(t))\right) > 0$ for any $t \in [0,1]$. Since $\boldsymbol{H}[F_\alpha](\boldsymbol{V}(0)) = \boldsymbol{I} > 0$ and the eigenvalues of $\boldsymbol{H}[F_\alpha](\boldsymbol{V}(t))$ change continuously w.r.t. $t \in [0,1]$, it results that $\boldsymbol{H}[F_\alpha](\boldsymbol{V}(1)) > 0$ for any $\boldsymbol{V}$, thus satisfying the condition of the theorem. ■

A fixed point $(\boldsymbol{\eta}^*, \boldsymbol{\lambda}^*)$ is stable if and only if $\max\{\rho(\boldsymbol{J}_{\boldsymbol{\eta}}(\boldsymbol{\lambda}^*(\boldsymbol{v}^*))), \rho(\boldsymbol{J}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*(\boldsymbol{v}^*)))\} < 1$. This implies $\sigma\left(\alpha^{-1}\mathrm{diag}(\boldsymbol{c}(\boldsymbol{V}^*)^2)\boldsymbol{\mathcal{M}}(\alpha)\right) \subseteq \mathbb{C}\setminus\mathbb{R}_{\geq 1}$ and leads to the following property.

Property 3. *Stable fixed points* $(\boldsymbol{\eta}^*, \boldsymbol{\lambda}^*)$ *of the damped Gaussian message passing algorithm* (2.18) *are local minima of the Gaussian Bethe free energy* F_α^c *in* (2.15) *at* $\boldsymbol{v}^*(\boldsymbol{\lambda}^*)$.

The above shows that the boundedness of F_α or the existence of local minima in case of an unbounded F_α plays a significant role in the convergence of Gaussian message passing. We illustrate this in Section 2.5. If the fractional message passing algorithm converges then it converges to a set of messages that corresponds to a local minimum of the fractional free energy. This also implies that the mean parameters of the local approximate marginals are exact (see Property 1. in Section 2.3). Note that the observations in Section 2.3 and Property A2 in the Appendix together with Property 3 imply that there is always a range of α values for which the fractional free energy possesses a local minimum to which the fractional message passing can converge.

2.4.3 The damping and the fractional parameter

The local stability condition in (2.24) is independent of the damping parameter ϵ. Therefore, it does not alter the local stability properties, it only makes the iteration slower and numerically more stable, that is, it dampens possible limit cycles caused by eigenvalues with real parts close to one.

The fractional parameter α characterizes the inference process and as we have seen in the example in the previous sections, by choosing smaller αs we can create local minima. There is a somewhat similar property for the message passing updates as well. Let $\Lambda \in \mathbb{R}^{|E|}$ be the set of messages $\boldsymbol{\lambda}$ that lead to normalizable approximate marginals. The set Λ is characterized by the model parameters $|\boldsymbol{R}|, \hat{\boldsymbol{\gamma}}$ and α. We reiterate that the elements of $\hat{\boldsymbol{v}}$ are the local variances v_{ij}^i and v_{ij}^j and there is a continuous bijective mapping between $\boldsymbol{\lambda} \in \Lambda$ and $\hat{\boldsymbol{v}} \in \mathbb{R}_+^{|E|}$ given by $\boldsymbol{y}(\hat{\boldsymbol{v}}) = \alpha\hat{\boldsymbol{\gamma}} + \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{\lambda}$, unless $\alpha = K$ and G is K-regular. This allows us to study the stability properties in terms of moment parameters $\hat{\boldsymbol{v}}(\boldsymbol{\lambda})$. Let $\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha) = v_{ij}(\alpha, v_{ij}^i, v_{ij}^j)/\sqrt{v_{ij}^i v_{ij}^j}$ be the vector of "local correlations". By using Gershgorin's theorem (Horn and Johnson, 2005) and $\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha)^2 \leq \boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha)$, we find that for any eigenvalue β of $\alpha^{-1}\mathrm{diag}(\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha))\boldsymbol{\mathcal{M}}(\alpha)$ or $\alpha^{-1}\mathrm{diag}(\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha))^2\boldsymbol{\mathcal{M}}(\alpha)$ we have

$$|\beta| \leq \max_{i,j} \left[\alpha^{-1}\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha)\left[(n_j - 1) + |1 - \alpha|\right]\right]. \tag{2.28}$$

Furthermore, $\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha)$ is an increasing function of α, and for any $\hat{\boldsymbol{v}} \in \mathbb{R}_+^{|E|}$

$$\lim_{\alpha \to 0} \alpha^{-1}\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha) = 0 \quad \text{and} \quad \lim_{\alpha \to \infty} \alpha^{-1}\boldsymbol{c}(\hat{\boldsymbol{v}}, \alpha) = 1. \tag{2.29}$$

These properties suggest that decreasing α might increase the chance of convergence, that is, the existence of a (locally) stable fixed point. However, due to the properties of $\boldsymbol{c}(\hat{\boldsymbol{v}})$, it is not easy to find a critical α that guarantees $|\beta| < 1$ for any $\hat{\boldsymbol{v}} \in \mathbb{R}^{|E|}$. This holds even when we restrict $\hat{\boldsymbol{v}}$ to the set of matched moment parameters $v_i = v_{ij}^i = v_{ik}^i$ with any $j, k \in \partial i$ and $v_i \in \mathbb{R}_+^n$.

2.5 Experiments

We implemented both direct minimization and fractional message passing and analyzed their behavior for different values of $\lambda_{max}(|\boldsymbol{R}|)$. For reasons of simplicity, we set all α_{ij}s equal. The results are summarized in Figure 2.3. Note that there is a good correspondence between the behavior of the fractional Bethe free energies in the direction of the eigenvalue corresponding to $\lambda_{max}(|\boldsymbol{R}|)$ and the convergence of the Newton method. The Newton method was started from different initial points. We experienced that when $\lambda_{max}(|\boldsymbol{R}|) > 1$ and setting the initial value to $\boldsymbol{v}_0 = t^2\boldsymbol{u}_{max}^2$, the algorithm did not converge for high values of t. This can be explained by the top plots in Figure 2.3: for high values of t, the initial point might not be in the convergence region of the local minimum. For the fractional message passing algorithm we used two types of initialization: (1) when $\lambda_{max}(|\boldsymbol{R}|) < 1$ we set Ψ_{ij} such that they are all normalizable by setting $\gamma_{ij}^i = |R_{ij}|u_{max}^j/\lambda_{max}u_{max}^i$ (e.g. Malioutov et al., 2006), (2) when $\lambda_{max}(|\boldsymbol{R}|) \geq 1$,

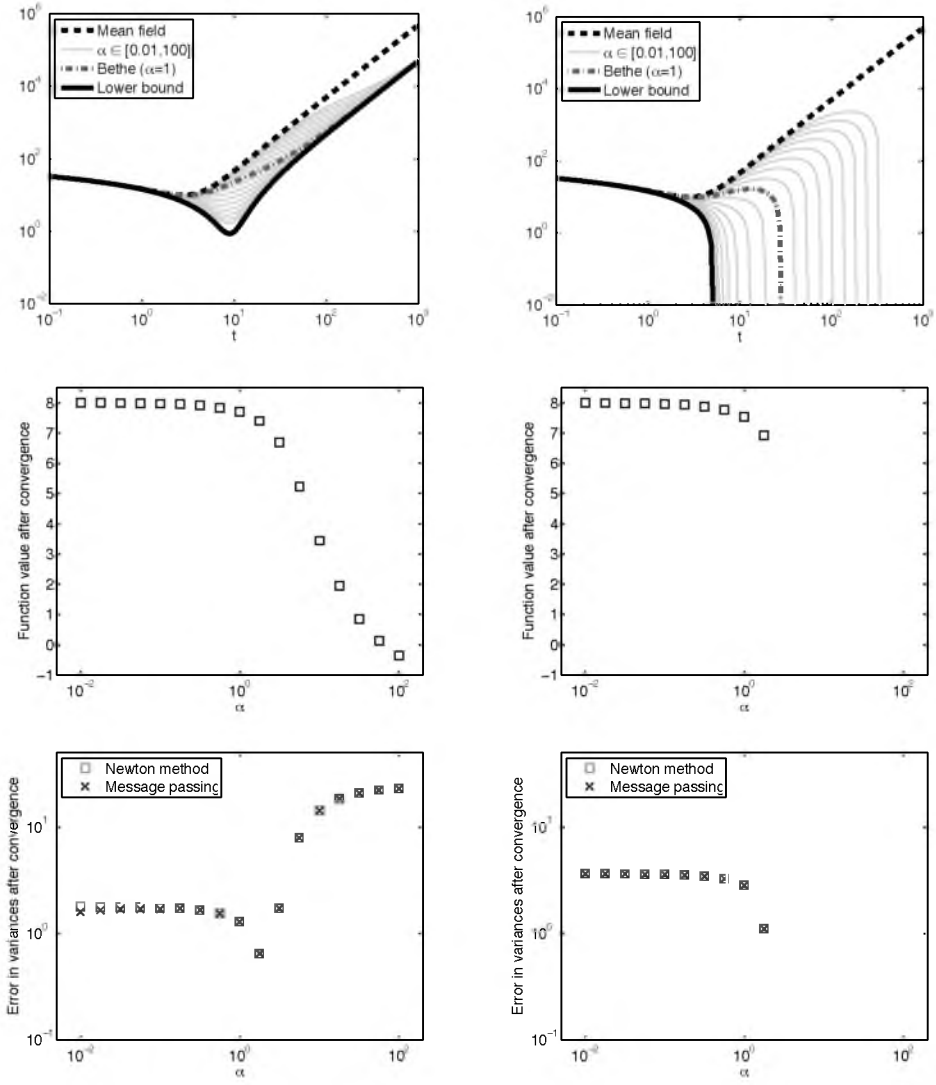

Figure 2.3: The top panels show the constrained fractional Bethe free energies of an 8 node Gaussian network in the direction $\sqrt{\boldsymbol{v}} = t\boldsymbol{u}_{max}$, where $\boldsymbol{u}_{max}$ is the eigenvector corresponding to $\lambda_{max}(|\boldsymbol{R}|)$ for $\lambda_{max}(|\boldsymbol{R}|) = 0.9$ (top-left) and $\lambda_{max}(|\boldsymbol{R}|) = 1.1$ (top-right). The thick lines are the functions F_{MF} (dashed), F_B (dashed dotted) and the lower bound $F_{\mathrm{MF}} - \frac{1}{2}\sqrt{\boldsymbol{v}}^T|\boldsymbol{R}|\sqrt{\boldsymbol{v}}$ (continuous). The thin lines are the constrained α-fractional free energies F_α^c for $\alpha \in [10^{-2}, 10^2]$. Center panels show the final function values after the convergence of the Newton method. The bottom panels show the $||\cdot||_2$ error in approximation for the single node standard deviations $\boldsymbol{\sigma} = \sqrt{\boldsymbol{v}}$. Missing values indicate non-convergence.

we used $\gamma_{ij}^i = 1/n_i$, that is, a symmetric partitioning of the diagonal elements. We set the initial messages such that all approximate marginals are normalizable in the first step of the iteration.

We experienced a behavior similar to that described by Welling and Teh (2001) for standard message passing, namely fractional message passing and direct minimization either both converge or both fail to converge. Our experiments in combination with Theorem 1 show that when $\lambda_{max}(|\boldsymbol{R}|) > 1$, standard message passing at best converges to a local minimum of the Bethe free energy. If standard message passing fails to converge, one can decrease α and search for a stationary point—preferably a local minimum—of the corresponding fractional free energy.

It can be seen from the results in the right panels of Figure 2.2, that when the model is no longer pairwise normalizable, the local minimum and not the unbounded global minimum can be viewed the natural continuation of the (bounded) global minimum for pairwise normalizable models. This explains why the quality of the approximation at the local minimum for models that are not pairwise normalizable is still comparable to that at the global minimum for models that are pairwise normalizable.

2.6 Conclusions

As we have seen, F_{MF} and $F_{\mathrm{MF}} - \frac{1}{2}\sqrt{\boldsymbol{v}}^T|\boldsymbol{R}|\sqrt{\boldsymbol{v}}$ provide tight upper and lower bounds for the Gaussian fractional Bethe free energies. It turns out that pairwise normalizability (see Malioutov et al., 2006) is not only a sufficient condition for the message passing algorithm to converge, but it is also a necessary condition for the Gaussian fractional Bethe free energies to be bounded from below.

If the model is pairwise normalizable, then the lower bound is bounded, and both direct minimization and message passing are converging. In our experiments both converged to the same minimum. This suggests that in the pairwise normalizable case, fractional Bethe free energies possess a unique global minimum.

If the model is not pairwise normalizable, then none of the fractional Bethe free energies are bounded from below. However, there is always a range of α values for which the fractional free energy possesses a local minimum to which both direct minimization and fractional message passing can converge. Thus, by decreasing α towards zero, one gets closer to the mean field energy and a finite local minimum will appear (Property A2 in the Appendix). We experienced that for a suitable range of αs, damping ϵ and initialization the fractional Gaussian message passing always converges.

As mentioned in Section 2.2.1, α_{ij}s correspond to using local α_{ij} divergences when applying power expectation propagation with a fully factorized approximating distribution. Seeger (2008) reports that when expectation propagation does not converge, applying power expectation propagation with $\alpha < 1$ helps to achieve convergence. In the case of the problem addressed in this chapter this behavior can be explained by the observation that small αs make a finite local minima more likely to occur and thus prevents the covariance matrices from becoming indefinite or even non positive definite. Although the most common reason for using $\alpha < 1$ in EP is numerical robustness, it also implies finding the saddle point of the α-fractional EP free energy. It might be interesting to investigate whether it is the same reason why convergence is more likely as in the case of Gaussian

fractional message passing.

Wainwright et al. (2003) propose to convexify the Bethe free energy for discrete models by choosing α_{ij}s sufficiently large such that the fractional Bethe free energy has a unique global minimum. This strategy appears to fail for Gaussian models. Convexification makes the possibly useful finite local minima disappear, leaving just the unbounded global minimum. In the case of the more general hybrid models, the use of the convexification is still unclear.

The example in Section 2.3 disproves the conjecture in Welling and Teh (2001): even when the Bethe free energy is not bounded from below, it can possess a finite local minimum to which the message passing and the minimization algorithms can converge.

We have shown that stable fixed points of the Gaussian fractional message passing algorithms are local minima of the fractional Bethe free energy. Although the existence of a local minimum does not guarantee the convergence of the message passing algorithm, in practice we experienced that the existence of a local minimum implies convergence. Based on these results, we *hypothesize* that when pairwise normalizability does not hold, the Gaussian Bethe free energy and the Gaussian message passing algorithm ($\alpha = 1$) can have two types of behavior:

(1) the Gaussian Bethe free energy possesses a unique finite local minimum to which optimization methods can converge by starting from, say, the mean field solution $v_i = 1/Q_{ii}$; the Gaussian message passing has a corresponding unique stable fixed point, to which it can converge with suitable starting point and sufficient damping,

(2) no finite local minimum exists, and thus, both the optimization and the message passing algorithm diverge.

By using the fractional free energy and the fractional message passing and by varying α, one can switch between these behaviors. Computing the critical $\alpha_c\left(|\boldsymbol{R}|\right)$ for a general $|\boldsymbol{R}|$ remains an open question. We believe that the properties of the free energies in K-regular symmetric models (Section 2.3), where the critical values can be easily computed, give a good insight into the properties of the free energies for general Gaussian models.

Chapter 3

Approximating marginals in latent Gaussian models

Summary

We consider the problem of correcting the Gaussian approximate posterior marginals computed by expectation propagation and the Laplace method in latent Gaussian models and propose correction methods that are similar in spirit to the Laplace approximation of Tierney and Kadane (1986). We show that in the case of sparse Gaussian models, the computational complexity of expectation propagation can be made comparable to that of the Laplace method by using a parallel updating scheme. In some cases, expectation propagation gives excellent estimates where the Laplace approximation fails. Inspired by bounds on the marginal corrections, we arrive at factorized approximations, which can be applied on top of both expectation propagation and the Laplace method. The factorized approximations give nearly indistinguishable results from the non-factorized approximations in a fraction of the time. This chapter is based on the material presented in Cseke and Heskes (2010a)[1] and it contains the results reported in Cseke and Heskes (2010c).

3.1 Introduction

Following Rue et al. (2009), we consider the problem of computing marginal probabilities over single variables in (sparse) latent Gaussian models. Probabilistic models with latent Gaussian variables are of interest in many areas of statistics, such as spatial data analysis (Rue and Held, 2005), and machine learning, such as Gaussian process models (e.g. Kuss and Rasmussen, 2005). The general setting considered in this chapter is as follows: the prior distribution over the latent variables is a Gaussian random field with a sparse precision (inverse covariance) matrix and the likelihood factorizes into a product of terms depending on just a single latent variable. Both the prior and the likelihood may

[1] B. Cseke and T. Heskes, *Improving posterior marginal approximations in latent Gaussian models*, AISTAS-2010, pages 121–128.

depend on a small set of hyper-parameters. We are interested in the posterior marginal probabilities over single variables given all observations.

Rue et al. (2009) propose an integrated nested Laplace approximation to approximate these posterior marginal distributions. Their procedure consists of three steps. 1) Approximate the posterior of the hyper-parameters given the data and use this to determine a grid of hyper-parameter values. 2) Approximate the posterior marginal distributions given the data and the hyper-parameters values on the grid. 3) Numerically integrate the product of the two approximations to obtain the posterior marginals of interest. The crucial contribution is the improved marginal posterior approximation in step 2), based on the approach of Tierney and Kadane (1986), that goes beyond the Gaussian approximation and takes into account higher order characteristics of (all) likelihood terms. Comparing their approach with Monte Carlo sampling techniques on several high-dimensional models, they show that their procedure is remarkably fast and accurate.

The main objective of the current chapter is to see whether we can improve upon the approach of Rue et al. (2009). Expectation propagation (Minka, 2001), a method for approximate inference developed and studied mainly in the machine learning community, is then an obvious candidate. It is well-known to yield approximations that are more accurate than the Laplace method (e.g. Minka, 2001; Kuss and Rasmussen, 2005). Furthermore, expectation propagation can still be applied in cases where the Laplace method is out of the question, for example, when the log-posterior is not twice-differentiable (Seeger, 2008). The typical price to be paid is that of higher computational complexity. However, we will see that, using a parallel instead of a sequential updating scheme, expectation propagation is at most a small constant factor slower than the Laplace method in applications on sparse Gaussian models with many latent variables. Moreover, along the way we will arrive at further approximations (both for expectation propagation and the Laplace method) that yield an order of magnitude speed-up, with hardly any degradation in performance.

The chapter is structured as follows. In Sections 3.1.1 and 3.2 we specify the model and briefly present the Laplace method and expectation propagation. In Section 3.3, we introduce and compare several methods for correcting marginals given a fixed setting of the hyper-parameters. In Section 3.4.6, we discuss the computational complexity of these methods when applied to sparse models. In Section 3.5, we introduce a method for numerical integration over hyper-parameters and finally in Section 3.6, we show that the proposed methods are competitive both in computational complexity and accuracy with the methods introduced in Rue et al. (2009). In order to increase the readability of the paper we include a schematic figure (Figure 3.13) and an explanatory list (Section 3.8) of the marginal approximation methods we introduce or refer to.

3.1.1 Latent Gaussian models

In this section, we introduce notation and define the model under consideration. Let $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_l)$ be the conditional probability of the observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ given the latent variables $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ and the hyper-parameters $\boldsymbol{\theta}_l$. We assume that the

likelihood $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_l)$ factorizes as

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_l) = \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}_l).$$

The prior $p(\boldsymbol{x}|\boldsymbol{\theta}_p)$ over the latent variables is taken to be Gaussian with canonical parameters $\boldsymbol{h}(\boldsymbol{\theta}_p)$ and $\boldsymbol{Q}(\boldsymbol{\theta}_p)$, that is,

$$p(\boldsymbol{x}|\boldsymbol{\theta}_p) \propto \exp\left(\boldsymbol{x}^T \boldsymbol{h}(\boldsymbol{\theta}_p) - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta}_p)\, \boldsymbol{x}\right).$$

Examples for $p(\boldsymbol{x}|\boldsymbol{\theta}_p)$ include Gaussian process models, where $\boldsymbol{Q}^{-1}(\boldsymbol{\theta}_p)$ is the covariance matrix at the input locations and Gaussian Markov random fields, where the elements of $\boldsymbol{Q}(\boldsymbol{\theta}_p)$ are the interactions strengths $Q_{ij}(\boldsymbol{\theta}_p)$ between the latent variables x_i and x_j. The prior $p(\boldsymbol{\theta}_l, \boldsymbol{\theta}_p)$ over the hyper-parameters is typically taken to be non-informative—uniform for location variables and log-uniform for scale variables—and factorizes w.r.t. $\boldsymbol{\theta}_l$ and $\boldsymbol{\theta}_p$. In order to simplify the notation, we use the proxy $\boldsymbol{\theta} = (\boldsymbol{\theta}_l, \boldsymbol{\theta}_p)$ to denote the hyper-parameters of the model.

The joint distribution of the variables in the model we study is

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}) \exp\left(\boldsymbol{x}^T \boldsymbol{h}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\, \boldsymbol{x}\right) p(\boldsymbol{\theta}).$$

We take $\boldsymbol{y}$ fixed and we consider the problem of computing accurate approximations of the posterior marginal densities of the latent variables $p(x_i|\boldsymbol{y}, \boldsymbol{\theta})$, given a fixed hyper-parameter value. Then we integrate these marginals over the approximations of the hyper-parameters posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$. The exact quantities are given by the formulas

$$p(x_i|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{1}{p(\boldsymbol{y}|\boldsymbol{\theta})} p(y_i|x_i, \boldsymbol{\theta}) \int d\boldsymbol{x}_{\setminus i}\, p(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{j \neq i} p(y_j|x_j, \boldsymbol{\theta}) \tag{3.1}$$

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{\theta}). \tag{3.2}$$

We use the term *evidence* for $p(\boldsymbol{y}|\boldsymbol{\theta}) = \int d\boldsymbol{x} p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$. In the following we omit $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$'s and $p(\boldsymbol{x}|\boldsymbol{\theta})$'s dependence on $\boldsymbol{\theta}$ whenever it is not relevant and use $t_i(x_i)$ as an alias of $p(y_i|x_i, \boldsymbol{\theta})$ and $p_0(\boldsymbol{x})$ as an alias of $p(\boldsymbol{x}|\boldsymbol{\theta})$. We use the notation $p(\boldsymbol{x}) = Z_p^{-1} p_0(\boldsymbol{x}) \prod_i t_i(x_i)$, with $Z_p(\boldsymbol{\theta}) \equiv p(\boldsymbol{y}|\boldsymbol{\theta})$. A Gaussian approximation of p will be denoted by q and Z_q will denote its normalization constant.

3.2 Global Gaussian approximations

A close inspection of (3.1) and (3.2) shows that computing $p(x_i|\boldsymbol{y}, \boldsymbol{\theta})$ boils down to computing similar integrals as for $p(\boldsymbol{y}|\boldsymbol{\theta})$. In this section, we review two approximation schemes that approximate such integrals: the Laplace method and expectation propagation (Minka, 2001). There are other approximation schemes, such as the variational approximation (e.g. Opper and Archambeau, 2009). The marginal approximation methods we propose for expectation propagation in Section 3.3 can be, under mild conditions,

translated to the variational approximation in Opper and Archambeau (2009). For this reason, we will not discuss the details of this method.

3.2.1 The Laplace method

The Laplace method approximates the evidence Z_p and, as a side product, it provides Gaussian approximation that is characterized by the local properties of the distribution at its mode $\boldsymbol{x}^* = \text{argmax}_{\boldsymbol{x}} \log p(\boldsymbol{x})$. The mean parameter of the corresponding approximating Gaussian density is $\boldsymbol{m} = \boldsymbol{x}^*$ while the inverse of the covariance parameter $\boldsymbol{V}$ is the Hessian of $-\log p$ at $\boldsymbol{x}^*$.

The idea behind the method is the following. Let $f = \log p$. Expanding f in second order at an arbitrary value $\tilde{\boldsymbol{x}}$, we get

$$\begin{aligned} f(\boldsymbol{x}) &= f(\tilde{\boldsymbol{x}}) + (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}) \\ &\quad + \frac{1}{2}(\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla^2_{\boldsymbol{xx}} f(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \tilde{\boldsymbol{x}}) + R_2[f](\boldsymbol{x}; \tilde{\boldsymbol{x}}), \end{aligned} \tag{3.3}$$

where $R_2[f](\boldsymbol{x}; \tilde{\boldsymbol{x}})$ is the residual term of the expansion at $\tilde{\boldsymbol{x}}$ with $R_2[f](\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{x}}) = 0$. By using the change of variables $\boldsymbol{s} = \boldsymbol{x} - \tilde{\boldsymbol{x}}$, we have

$$\begin{aligned} \log \int d\boldsymbol{x} e^{f(\boldsymbol{x})} &= f(\tilde{\boldsymbol{x}}) - \frac{1}{2}\nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}})^T \left[\nabla^2_{\boldsymbol{xx}} f(\tilde{\boldsymbol{x}})\right]^{-1} \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}) \\ &\quad - \frac{1}{2}\log|-\nabla^2_{\boldsymbol{xx}} f(\tilde{\boldsymbol{x}})| + \log \mathrm{E}_{\boldsymbol{s}}\left[e^{R_2[f](\boldsymbol{s}+\tilde{\boldsymbol{x}};\tilde{\boldsymbol{x}})}\right], \end{aligned} \tag{3.4}$$

where $|\cdot|$ denotes the determinant and the expectation w.r.t. $\boldsymbol{s}$ is taken over a normal distribution with canonical parameters $\nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}})$ and $-\nabla^2_{\boldsymbol{xx}} f(\tilde{\boldsymbol{x}})$.

A closer look at (3.3) and (3.4) suggests that choosing $\tilde{\boldsymbol{x}} = \boldsymbol{x}^*$ and using the approximation $R_2[\log p](\boldsymbol{x}; \tilde{\boldsymbol{x}}) \approx 0$ yields an approximation of the log evidence

$$\log \int d\boldsymbol{x} p(\boldsymbol{x}) \approx \log p(\boldsymbol{x}^*) - \frac{1}{2}\log|-\nabla^2_{\boldsymbol{xx}} \log p(\boldsymbol{x}^*)|. \tag{3.5}$$

Meanwhile, p can be approximated by the Gaussian

$$q(\boldsymbol{x}) = N\left(\boldsymbol{x}|\boldsymbol{x}^*, -\left[\nabla^2_{\boldsymbol{xx}} \log p(\boldsymbol{x}^*)\right]^{-1}\right). \tag{3.6}$$

Note that any reasonably good approximation of $\mathrm{E}_{\boldsymbol{s}}\left[e^{R_2[f](\boldsymbol{s}+\tilde{\boldsymbol{x}};\tilde{\boldsymbol{x}})}\right]$ can improve the accuracy of the approximation in (3.5).

The Laplace method requires the second order differentiability of $\log p$ at $\boldsymbol{x}^*$, thus a sufficient condition for the applicability of this approximation scheme is the second order differentiability of $\log p$. The necessary condition is the second order differentiability at the mode $\boldsymbol{x}^*$. A distribution p for which the method fails to give any meaningful information about the variances is, for example, when $p(y_j|x_j) = \lambda \exp(-\lambda|y_j - x_j|)/2$. In this case, the Hessian of $\log p$ at an arbitrary point $\tilde{\boldsymbol{x}}$ is either equal to the precision $\boldsymbol{Q}$ of the prior or it is undefined. Since the Laplace method captures the characteristics of the modal configuration, it often gives poor estimates of the normalization constant (e.g. Kuss and Rasmussen, 2005). The example in Section 3.4.1 shows how this behavior influences

the approximation of the marginals in case of a two dimensional toy model. However, compared to other methods, the main advantage of the Laplace method is its speed. The optimization of $\log p$ w.r.t. $\boldsymbol{x}$ for computing $\boldsymbol{m} = \boldsymbol{x}^*$ requires only a few Newton steps.

3.2.2 Expectation propagation

Expectation propagation (EP) approximates the integral for computing the evidence in the following way. Let us assume that q is a Gaussian approximation of p constrained to have the form $q(\boldsymbol{x}) = Z_q^{-1} p_0(\boldsymbol{x}) \prod_j \tilde{t}_j(x_j)$. Then the evidence can be approximated as

$$\begin{aligned} Z_p &= \int d\boldsymbol{x}\, p_0(\boldsymbol{x}) \prod_j t_j(x_j) \\ &= Z_q \int d\boldsymbol{x}\, q(\boldsymbol{x}) \prod_j \frac{t_j(x_j)}{\tilde{t}_j(x_j)} \\ &\approx Z_q \prod_j \int dx_j\, q(x_j) \frac{t_j(x_j)}{\tilde{t}_j(x_j)}. \end{aligned} \tag{3.7}$$

and we are left with choosing the appropriate $\tilde{t}_j(x_j)$s that yield both a good approximation of the evidence and of $p(\boldsymbol{x})$. EP computes the terms $\tilde{t}_j(x_j)$ by iterating

$$\tilde{t}_j^{new}(x_j) \propto \frac{\text{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1} q(\boldsymbol{x})\right)}{q(\boldsymbol{x})} \tilde{t}_j(x_j), \text{ for all } j = 1, \ldots, n, \tag{3.8}$$

where $\text{Collapse}(r) = \text{argmin}_{r' \in \mathcal{N}} \mathrm{D}\left[r \,||\, r'\right]$ is the Kullback-Leibler (KL) projection of the distribution r into the family of Gaussian distributions $\mathcal{N}$. In other words, it is the Gaussian distribution that matches the first two moments of r. Using the properties of the KL divergence, one can check that when the terms t_j depend only on the variables x_j then $\text{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1} q(\boldsymbol{x})\right)/q(\boldsymbol{x}) = \text{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1} q(x_j)\right)/q(x_j)$, therefore, the iteration in (3.8) is well defined. At any fixed point of this iteration, we have a set of $\tilde{t}_j(x_j)$ terms for which $\text{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1} q(\boldsymbol{x})\right) = q(\boldsymbol{x})$ for any $j \in \{1, \ldots, n\}$. By defining the cavity distribution $q^{\setminus j}(\boldsymbol{x}) \propto \tilde{t}_j(x_j)^{-1} q(\boldsymbol{x})$ and scaling the terms t_j, the above fixed point condition can be rewritten as

$$\int dx_j \left\{1, x_j, x_j^2\right\} q^{\setminus j}(x_j)\tilde{t}_j(x_j) = \int dx_j \left\{1, x_j, x_j^2\right\} q^{\setminus j}(x_j) t_j(x_j),\ j = 1, \ldots, n,$$

and so, the approximation for Z_p has the form

$$Z_p \approx \int d\boldsymbol{x}\, p_0(\boldsymbol{x}) \prod_j \tilde{t}_j(x_j).$$

Expectation propagation, can be viewed as a generalization of loopy belief propagation (e.g. Murphy et al., 1999) to probabilistic models with continuous variables and also as an iterative application of the assumed density filtering procedure (e.g. Csató and Opper, 2001). A close inspection of the parametric form of the iteration in Section A.4 of

the Appendix shows that the convexity of $\log \int d\boldsymbol{x}\, N(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{V})\, t_j(x_j)$ w.r.t. $\boldsymbol{m}$ or the concavity of $\log t_j(x_j)$ (Seeger, 2008) is a sufficient condition for the terms $\tilde{t}_j$s to be normalizable and thus for the existence of q^{new}. However, this alone does not guarantee convergence. To our knowledge, the issue of EP's convergence in case of the models we study in this chapter is still an open question. The iteration in (3.8) can also be derived by using variational free energies (e.g. Heskes et al., 2005; Minka, 2005). It can be relaxed such that the projections are taken on $t_j(x_j)^{\alpha}\tilde{t}_j(x_j)^{-\alpha}q(\boldsymbol{x})$, with $\alpha \in (0, 1]$. The limit $\alpha \to 0$ corresponds to the variational approximation of Opper and Archambeau (2009).

When applying EP to models with Gaussian Markov random field priors it is often desirable to be able to deal with (deterministic) linear constraints of the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ (e.g. Rue et al., 2009). To incorporate these constraints into EP would require to deal with terms of the form $\delta_0(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$. These types of terms require a special treatment. In the following we derive a possible way to deal with them. First we start out by deriving a sampling distributions for the Gaussian random variables $\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where we assume that $\boldsymbol{A}$ is a $k \times n$ matrix with $k < n$. Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{V})$ and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, v\boldsymbol{I})$. Then the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ is a Gaussian with parameters $\boldsymbol{m} + \boldsymbol{V}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T + v\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{m} + \boldsymbol{b})$ and $\boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T + v\boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{V}$. Setting $\boldsymbol{y} = 0$ and taking the limit $v \to 0$ we find that

$$\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{m} - \boldsymbol{V}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{b}), \boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{V}\right). \quad (3.9)$$

As a consequence we propose the following strategy to deal with linear constraint in EP: (1) we perform term updates on all "regular" terms and before starting a new update we project the new parameters of q according to (3.9), (2) the value of the corresponding factor in (3.7) is $N(\boldsymbol{0}|\boldsymbol{A}\boldsymbol{m} - \boldsymbol{b}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T)$ and it corresponds to a Bayesian update in the limit $v \to 0$.

3.3 Approximation of posterior marginals

The global approximations provide Gaussian approximations q of p and approximations of the evidence Z_p. The Gaussian approximation q can be used to compute Gaussian approximations of posterior marginals. In case of the Laplace method this only requires linear algebraic methods (computing the diagonal elements of the Hessian's inverse), while in the case of EP, the approximate marginals are a side product of the method itself. We refer to the corresponding Gaussian marginal approximations by LM (Laplace method) and EP (EP). Moreover, one can make use of the approximation method at hand in order to improve the Gaussian approximate marginals.

In case of the Laplace method, one can easily check that the residual term in (3.3) decomposes as $R_2[\log p](\boldsymbol{x}; \tilde{\boldsymbol{x}}) = \sum_j R_2[\log t_j](x_j; \tilde{x}_j)$, thus, when approximating the marginal of x_i it is sufficient to assume $R_2[\log t_j](x_j; \tilde{x}_j) \approx 0$ only for $j \neq i$. This yields a locally improved approximation $q(x_i) \times \exp R_2[\log t_i](x_i; x_i^*)$ to which we refer by LM-L.

As shown in Section 3.2.2, EP is basically built on exploiting the low–dimensionality of $t_i(x_i)$ and approximating the *tilted* marginals $t_i(x_i)q^{\backslash i}(x_i)$. These are known to be better approximations of the marginals $p(x_i)$ than $q(x_i)$ (e.g. Opper and Winther, 2000; Opper et al., 2009). We refer to this approximation by EP-L.

These observations show that there are ways to improve the marginals of the global approximation q by exploiting the properties of the methods. For the moment, however, we postpone this to Section 3.4 and first try to compute the marginals from scratch. This gives us some insight into where to look for further improvements.

The exact marginals can be computed as

$$p(x_i) = \frac{1}{Z_p} t_i(x_i) \int d\boldsymbol{x}_{\setminus i}\, p_0(\boldsymbol{x}_{\setminus i}, x_i) \prod_{j \neq i} t_j(x_j), \tag{3.10}$$

thus, as mentioned earlier, computing the marginal for a fixed x_i boils down to computing the normalization constant of the distribution $p_0(\boldsymbol{x}_{\setminus i}|x_i) \prod_{j\neq i} t_j(x_j)$. Therefore, we can use our favorite method to approximate it. In the following, we present the details of these procedures for the Laplace method and EP.

3.3.1 Laplace approximation

We use the same line of argument as in Section 3.2.1, but now we fix x_i and expand $\log p$ w.r.t. $\boldsymbol{x}_{\setminus i}$ at an arbitrary $\tilde{\boldsymbol{x}}_{\setminus i}$. The expression is identical to (3.3) with $\boldsymbol{x} = (x_i, \boldsymbol{x}_{\setminus i}^T)^T$ and $\tilde{\boldsymbol{x}} = (x_i, \tilde{\boldsymbol{x}}_{\setminus i}^T)^T$. Let $\boldsymbol{x}_{\setminus i}^*(x_i) = \operatorname{argmax}_{\boldsymbol{x}_{\setminus i}} \log p(x_i, \boldsymbol{x}_{\setminus i})$ and let $\tilde{\boldsymbol{x}}_{\setminus i} = \boldsymbol{x}_{\setminus i}^*(x_i)$. Then the approximation of (3.4) simplifies to a form similar to (3.5), that is, the approximation of the marginal density, up to the constant $\log Z_p$, is given by

$$\log \int d\boldsymbol{x}_{\setminus i} p(\boldsymbol{x}) \approx \log p\left(x_i, \boldsymbol{x}_{\setminus i}^*(x_i)\right) - \frac{1}{2} \log \left| -\nabla^2_{\boldsymbol{x}_{\setminus i}\boldsymbol{x}_{\setminus i}} \log p\left(x_i, \boldsymbol{x}_{\setminus i}^*(x_i)\right) \right|. \tag{3.11}$$

This approximation is known in statistics as the Laplace approximation (Tierney and Kadane, 1986) and we will refer to it as $\tilde{p}_i^{\text{LA-TK}}(x_i)$.

The error of the approximation can be characterized in terms of the residual terms of the second order expansion. The residual decomposes as

$$R_2[\log p](\boldsymbol{x}; \tilde{\boldsymbol{x}}) = \sum_{j \neq i} R_2[\log t_j]\left(s_j + x_j^*(x_i); x_j^*(x_i)\right)$$

and the expectation is taken w.r.t. $\boldsymbol{s} \in \mathbb{R}^{(n-1)}$ having a normal density with mean 0 and inverse covariance $-\nabla^2_{\boldsymbol{x}_{\setminus i}\boldsymbol{x}_{\setminus i}} \log p(x_i, \boldsymbol{x}_{\setminus i}^*(x_i))$. This means that in principle we have exact estimates of the error and that any reasonable approximation of the integral can improve the quality of the approximation in (3.11).

3.3.2 Expectation propagation

The integral in (3.10) can also be approximated using EP. As mentioned above EP typically provides better approximations of $\log Z_p$ than the Laplace method. For this reason, the marginals computed by approximating (3.10) using EP are expected to be more accurate. The procedure is as follows: (1) fix x_i and compute the canonical parameters of $p_0(\boldsymbol{x}_{\setminus i}|x_i)$ given by $\boldsymbol{h}_{\setminus i} - \boldsymbol{Q}_{\setminus i,i} x_i$ and $\boldsymbol{Q}_{\setminus i,\setminus i}$ and (2) use EP to approximate the integral in (3.10). Thus we approximate the integral by leaving out $p_0(x_i)$ and $t_i(x_i)$ and applying EP using the prior $p_0(\boldsymbol{x}_{\setminus i}|x_i)$ and the terms $t_j(x_j)$, $j \neq i$.

3.4 Approximation of posterior marginals by correcting the global approximations

As we have seen in the previous section, computing the marginal for a given fixed x_i value can be as expensive as the global procedure itself. On the other hand, however, there are ways to improve the marginals of the global approximation. In this section, we start from the "direct" approach and try to re-use the results of the global approximation to improve on the locally improved marginals LM-L and EP-L.

We start with the observation that for all the presented approximation methods, we can write the approximating distribution q as

$$q(\boldsymbol{x}) = \frac{1}{Z_q} p_0(\boldsymbol{x}) \prod_j \tilde{t}_j(x_j). \tag{3.12}$$

In case of the Laplace method, the canonical parameters of the Gaussian functions $\tilde{t}_j$ are defined by the parameters of the Taylor expansion of $\log t_j$ at x_i^*, while in case of EP, they are the parameters corresponding to EP's fixed point.

In the following, we do not keep track of the normalization constants that are independent of x_i. In order to avoid overloading the notation and to express that a distribution is approximated as proportional to an expression on the right hand side of the $\approx$ relation, we occasionally use Z as a proxy for unknown normalization constants. One can keep track of these constants, but in most cases, from the practical point of view, it is easier to perform a univariate numerical interpolation followed by numerical quadrature and (re)normalization.

3.4.1 Marginal corrections

Given a global Gaussian approximation $q(\boldsymbol{x})$ of the form (3.12) with corresponding term approximations $\tilde{t}_i(x_i)$, we can rewrite $p(x_i)$ as

$$\begin{aligned} p(x_i) &= \frac{Z_q}{Z_p} \frac{t_i(x_i)}{\tilde{t}_i(x_i)} \int d\boldsymbol{x}_{\setminus i}\, q(\boldsymbol{x}) \prod_{j \neq i} \frac{t_j(x_j)}{\tilde{t}_j(x_j)} \\ &= \frac{Z_q}{Z_p} \frac{t_i(x_i)}{\tilde{t}_i(x_i)} q(x_i) \int d\boldsymbol{x}_{\setminus i}\, q(\boldsymbol{x}_{\setminus i}|x_i) \prod_{j \neq i} \frac{t_j(x_j)}{\tilde{t}_j(x_j)} \\ &= \frac{Z_q}{Z_p} \epsilon_i(x_i) q(x_i) \int d\boldsymbol{x}_{\setminus i}\, q(\boldsymbol{x}_{\setminus i}|x_i) \prod_{j \neq i} \epsilon_j(x_j), \end{aligned} \tag{3.13}$$

where we define $\epsilon_i(x_i) \equiv t_i(x_i)/\tilde{t}_i(x_i)$. In case of EP, the term approximations $\tilde{t}_i(x_i)$ are chosen to be close to the terms $t_i(x_i)$ in average w.r.t. $q(x_i)$. For this reason, we expect the $\epsilon_i(x_i)$'s to be close to 1 in average w.r.t $q(x_i)$.

Equation (3.13) is still exact and it shows that there are two corrections to the Gaussian approximation $q(x_i)$: one direct, local correction through $\epsilon_i(x_i)$ and one more indirect correction through the (weighted integral over) $\epsilon_j(x_j)$s for $j \neq i$. The direct, local correction comes without additional cost and suggests the above–mentioned (Section 3.3)

local approximation

$$p(x_i) \approx \frac{1}{Z}\epsilon_i(x_i)q(x_i) .$$

We use the notations $\tilde{p}_i^{\text{EP-L}}(x_i)$ and $\tilde{p}_i^{\text{LM-L}}(x_i)$ for the approximations following the global Gaussian approximations by EP and Laplace method, respectively.

To improve upon this approximation, we somehow have to get a handle on the indirect correction

$$c_i(x_i) \equiv \int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}_{\setminus i}|x_i\right) \prod_{j\neq i} \epsilon_j(x_j) . \tag{3.14}$$

Again, for each x_i, we are in fact back to the form (3.10): we have to estimate the normalization constant of a latent Gaussian model, where $q\left(\boldsymbol{x}_{\setminus i}|x_i\right)$ now plays the role of an $(n-1)$-dimensional Gaussian prior and the $\epsilon_j(x_j)$s are terms depending on a single variable. Running a complete procedure, be it EP or Laplace, for each x_i—as described in Sections 3.3.1 and 3.3.2—is often computationally too intensive and further approximations are needed to reduce the computational burden.

EP corrections

Let us write $\tilde{\epsilon}_j(x_j; x_i)$ for the term approximation of $\epsilon_j(x_j)$ in the context of approximating $c_i(x_i)$. A full run of EP for each x_i may be too expensive, so instead we propose to perform just one simultaneous EP step for all $j \neq i$. Since the term approximations of the global EP approximation are tuned to make $\tilde{t}_j(x_j)$ close to $t_j(x_j)$ w.r.t. $q(x_i)$, it makes sense to initialize $\tilde{\epsilon}_j(x_j; x_i)$ to 1. Following EP, computing the new term approximation for term j then amounts to choosing $\tilde{\epsilon}_j(x_j; x_i)$ such that

$$\int dx_j\, \{1, x_j, x_j^2\} q(x_j|x_i) \tilde{\epsilon}_j(x_j; x_i) = \int dx_j\, \{1, x_j, x_j^2\} q(x_j|x_i) \epsilon_j(x_j), \tag{3.15}$$

that is, we get $\tilde{\epsilon}_j(x_j; x_i)$ by *collapsing* $\epsilon_j(x_j; x_i) q(x_j|x_i)$ into a Gaussian and dividing it by $q(x_j|x_i)$. As we have seen in Section 3.2.2, EP computes $\tilde{t}_j$ such that

$$\int dx_j\, \{1, x_j, x_j^2\} q(x_j) = \int dx_j\, \{1, x_j, x_j^2\} q(x_j) \epsilon_j(x_j), \tag{3.16}$$

thus, the difference here is made by the conditioning on x_i and $\tilde{\epsilon}_j(x_j; x_i)$ can be viewed as an update $\tilde{t}_j(x_j; x_i)$ of $\tilde{t}_j(x_j)$ that accounts "locally" for this difference—up to second order. Replacing the terms $\epsilon_j(x_j)$ in (3.14) by their term approximations $\tilde{\epsilon}_j(x_j; x_i)$ yields an estimate for $c_i(x_i)$. The corresponding approximation

$$p(x_i) \approx \frac{1}{Z}\epsilon_i(x_i)q(x_i) \int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}_{\setminus i}|x_i\right) \prod_{j\neq i} \tilde{\epsilon}_j(x_j; x_i) \tag{3.17}$$

is referred to as $\tilde{p}_i^{\text{EP-1STEP}}(x_i)$. By performing further EP steps, one can refine the term approximations $\tilde{\epsilon}_j(x_j; x_i)$. Iterating the EP steps until convergence (as mentioned above)

leads to a similar (costly) approximation as in Section 3.3.2. We refer to the resulting approximation as EP-FULL.

Laplace corrections

According to the Laplace approximation presented in Section 3.3.1 one has to recompute the conditional mode $\boldsymbol{x}^*_{\backslash i}(x_i)$ for every choice of x_i. In order to lessen the computational burden, Rue et al. (2009) propose to re-use the global approximation by approximating the conditional mode with the conditional mean, that is, $\boldsymbol{x}^*_{\backslash i}(x_i) \approx \boldsymbol{m}_{\backslash i} + \boldsymbol{V}_{\backslash i,i} V_{i,i}^{-1}(x_i - m_i)$, where $\boldsymbol{m} = \boldsymbol{x}^* (= \operatorname{argmax}_{\boldsymbol{x}} \log p(\boldsymbol{x}))$. This approximation often performs reasonably well when p is close to a Gaussian.

In our setting, the approximation proposed by Rue et al. (2009) can be understood as follows. The error terms ϵ_j can be identified as $\log \epsilon_i(x_i) = R_2 \left[\log t_i\right](x_i; m_i)$. In order to assess $c_i(x_i)$, one could, in principle, apply the Laplace method to

$$f(\boldsymbol{x}_{\backslash i}; x_i) \equiv q\left(\boldsymbol{x}_{\backslash i} | x_i\right) \prod_{j \neq i} \epsilon_j(x_j) \,.$$

This would be identical to the direct method of Tierney and Kadane (1986) presented in Section 3.3.1. Using the conditional mean as an approximation of the conditional mode boils down to ignoring the terms $\epsilon_j(x_j)$ and using the mode of $q(\boldsymbol{x}_{\backslash i}|x_i)$. The corresponding approximation is of the form (3.17), where now $\tilde{\epsilon}_j(x_j; x_i)$ follows from a second-order Taylor expansion of $\log \epsilon_j(x_j)$ around the mode or mean of $q(x_j|x_i)$. We refer to this approximation as $\tilde{p}_i^{\text{LA-CM}}(x_i)$.

Taking a closer look at (3.4) and using our assumptions in Section 3.3.1, we can easily see that when we are not evaluating the normalization constant at the conditional mode, we can refine the approximation by adding $-\frac{1}{2} \nabla_{\boldsymbol{x}_{\backslash i}} f(\tilde{\boldsymbol{x}}_{\backslash i}) [\nabla^2_{\boldsymbol{x}_{\backslash i}, \boldsymbol{x}_{\backslash i}} f(\tilde{\boldsymbol{x}}_{\backslash i})]^{-1} \nabla_{\boldsymbol{x}_{\backslash i}} f(\tilde{\boldsymbol{x}}_{\backslash i})$, which is not identical to zero when the expansion in not made at the mode, that is, when $\tilde{\boldsymbol{x}}_{\backslash i} \neq \boldsymbol{x}^*_{\backslash i}(x_i)$. As we will see in Section 3.4.7, this correction adds no significant computational burden to the method proposed in Rue et al. (2009). We refer to this approximation as $\tilde{p}_i^{\text{LA-CM2}}(x_i)$.

In order to further reduce computational effort, Rue et al. (2009) suggest additional approximations. Because they can only be expected to reduce the accuracy of the final approximation, we will not consider them in our experiments in Sections 3.4.5 and 3.6. Below we propose another EP–related approximation, motivated by theoretical bounds on the corrections $c_i(x_i)$.

3.4.2 Bounds and factorized approximations

The computational bottleneck in the above procedures for approximating the correction $c_i(x_i)$ is not computing appropriate approximations of the terms $\epsilon_j(x_j)$, either through EP or Laplace, but instead computing the normalization of the resulting Gaussian form in (3.17), which boils down to the computation of a Gaussian normalization constant. Here we propose a simplification, which we motivate through its connection to bounds on the marginal correction $c_i(x_i)$.

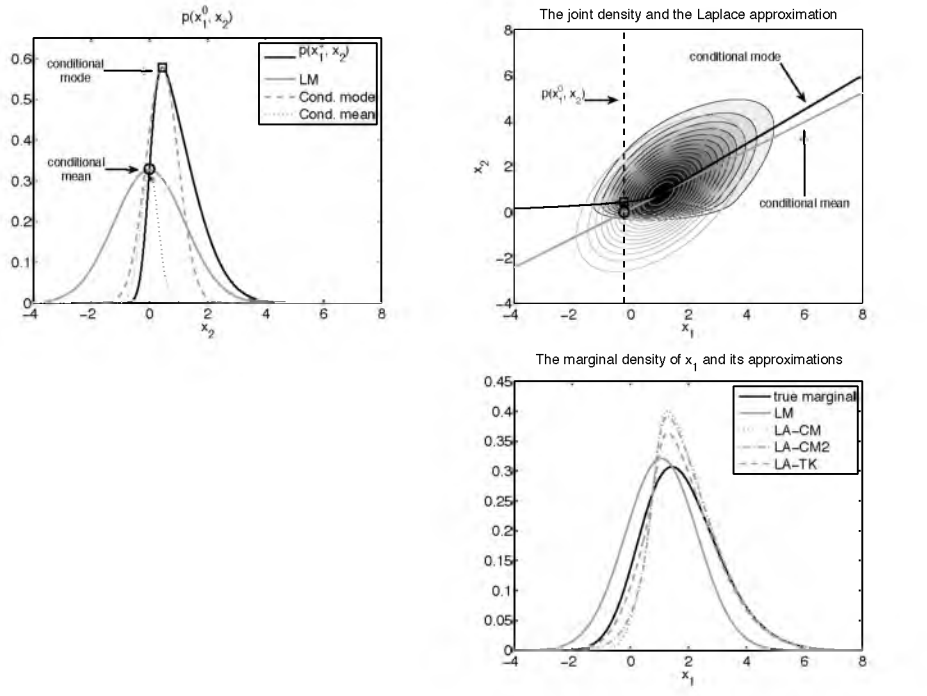

Figure 3.1: A two-dimensional example, illustrating how Laplace approximation works and why it can fail. On the top-right panel, the black contour curves show the true distribution, the gray contour curves stand for the global Laplace approximation, and the black and gray curves show the conditional modes and the conditional means w.r.t. x_1. The square and circle outline these quantities for a fixed x_1^0. The dashed vertical line emphasizes the "slice" $p(x_1^0, x_2)$ at x_1^0. The top-left panel shows $p(x_1^0, x_2)$ and the approximations for computing its area under the curve. The areas under the Gaussian curves corresponding to the conditional mode (square) and the conditional mean (circle) are the approximations of $p(x_1^0) = \int dx_2\, p(x_1^0, x_2)$. The bottom-right panel shows the marginal of $p(x_1)$ and its approximations. The conditional mean can severely underestimate the mass for $x_1 = x_1^0$.

Using Jensen's inequality, we obtain the lower bound on (3.14)

$$c_i(x_i) \geq \exp\left[\sum_{j\neq i}\int dx_j\, q(x_j|x_i)\log\epsilon_j(x_j)\right] \equiv c_i^{\text{lower}}(x_i)\,.$$

Following Minka (2005), we can also get an upper bound:

$$c_i(x_i) \leq \prod_{j\neq i}\left[\int dx_j\, q(x_j|x_i)\epsilon_j(x_j)^{n-1}\right]^{1/(n-1)} \equiv c_i^{\text{upper}}(x_i)\,.$$

This upper bound will in many cases be useless because the integral often does not exist. The lower bound, which corresponds to a mean-field-type approximation, does not have this problem, but may still be somewhat conservative. We therefore propose the general family of approximations

$$c_i^{(\alpha)}(x_i) = \prod_{j\neq i}\left[\int dx_j\, q(x_j|x_i)\epsilon_j(x_j)^{\alpha}\right]^{1/\alpha}\,. \tag{3.18}$$

It is easy to show that

$$c_i^{\text{lower}}(x_i) \leq c_i^{(\alpha)}(x_i) \leq c_i^{\text{upper}}(x_i) \quad \forall\, 0 \leq \alpha \leq n-1\,,$$

where $\alpha = 0$ is interpreted as the limit $\alpha \to 0$. Furthermore, for any α we obtain exactly the same Taylor expansion in terms of $\epsilon_j(x_j) - 1$ (see Opper et al. (2009) and Section 3.4.3 below). The choice $\alpha = 1$ makes the most sense: it gives exact results for $n = 2$ as well as when all x_js are indeed conditionally independent given x_i. We refer to the corresponding approximation as $\tilde{p}_i^{\text{EP-FACT}}(x_i)$.

Using (3.15), it is easy to see that $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ corresponds to $\tilde{p}_i^{\text{EP-1STEP}}(x_i)$ if in (3.17) we would replace $q(\boldsymbol{x}_{\setminus i}|x_i)$ by the factorization $\prod_{j\neq i} q(x_j|x_i)$, i.e., as if the variables x_j in the global Gaussian approximation are conditionally independent given x_i. The same replacement in the Laplace approximation yields the approximation referred to as $\tilde{p}_i^{\text{LA-FACT}}(x_i)$. Here, we compute the univariate integrals with the Laplace method and using the approximation $x_j^*(x_i) \approx \mathrm{E}_q[x_j|x_i]$, with $q(\boldsymbol{x})$ being the global approximation resulting from the Laplace method.

The factorization principle may be applied to groups of variables as well. We can use the idea in Section 3.4.1 for whole groups of variables $\boldsymbol{x}_I$ by factorizing $q(\boldsymbol{x}_{\setminus I}|\boldsymbol{x}_I)$. Another way to make use of this principle is by using it recursively. In this way, we can obtain higher order corrections of the approximate marginals and the evidence approximation. We will detail these methods in a future report.

One of the advantages of the bounding arguments is that we can extend the factorized approximation to cases when t_j depends on more variables, say, $\boldsymbol{x}_{I_j}$, with $I_j \in \{1,\ldots,n\}$. In this case, the factorization is unfeasible since $\prod_j t_j\left(\boldsymbol{x}_{I_j}\right)$ may not factorize w.r.t. x_j. By using the bounding argument (Minka, 2005), we can still compute

a "factorized" approximation

$$c_i^{(\alpha)}(x_i) = \prod_{j \neq i} \left[\int d\boldsymbol{x}_{I_j} \, q(\boldsymbol{x}_{I_j}|x_i) \epsilon_j(\boldsymbol{x}_{I_j})^\alpha \right]^{1/\alpha}.$$

A similar argument can be applied when t_i depends on a linear transformation of the variables, for example, in logistic regression models.

3.4.3 Connection to the Taylor expansion in Opper et al. (2009)

The line of argument in Opper et al. (2009) when applied to approximating the marginals can be explained in our notation as follows. By expanding $p(\boldsymbol{x}) = Z_q Z_p^{-1} q(\boldsymbol{x}) \prod_j \epsilon_j(x_j)$ in first order w.r.t. all $\epsilon_j(x_j) - 1$, they obtain a first order approximation of the exact p in terms of the global approximation q and the tilted distributions $t_j(x_j)q^{\backslash j}(\boldsymbol{x})$. The marginalization of this expansion yields the marginal approximation

$$\tilde{p}_i^{\text{EP-OPW}}(x_i) \equiv \frac{Z_q}{Z_p} q(x_i) \left[1 + \sum_j \int dx_j q(x_j|x_i)[\epsilon_j(x_j) - 1] \right].$$

Since the goal of Opper et al. (2009) was to provide improved approximations of the posterior distribution $p(\boldsymbol{x})$, and not only of its marginals, a natural adaptation of their approach would be to expand w.r.t. to all $j \neq i$ and not i itself. This leads to the approximation

$$p(x_i) \approx q(x_i)\epsilon_i(x_i) \left[1 + \sum_{j \neq i} \int dx_j q(x_j|x_i)[\epsilon_j(x_j) - 1] \right],$$

which is also the first order expansion of $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ w.r.t. $\epsilon_j(x_j) - 1$, $j \neq i$. A further expansion w.r.t $\epsilon_i(x_i) - 1$ leads to $\tilde{p}_i^{\text{EP-OPW}}(x_i)$, thus the two approximations are equal in first order. An advantage of $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ is that it is non-negative by construction, while $\tilde{p}_i^{\text{EP-OPW}}(x_i)$ can take on negative values.

3.4.4 Approximating predictive densities in Gaussian processes

In many real-world problems, the prior $p_0(\boldsymbol{x})$ is defined as a Gaussian process—most often in terms of moment parameters—and besides marginals, one is also interested in computing accurate approximations of the predictive densities

$$p(\boldsymbol{x}_*|\boldsymbol{y}) = Z_p^{-1} \int d\boldsymbol{x} \, p_0(\boldsymbol{x}_*|\boldsymbol{x}) p_0(\boldsymbol{x}) \prod_j t_j(x_j),$$

where $\boldsymbol{x}_*$ is a set of latent variables of which distribution we aim to approximate. By defining the $\hat{q}(\boldsymbol{x}, \boldsymbol{x}_*) \propto p_0(\boldsymbol{x}_*|\boldsymbol{x})q(\boldsymbol{x})$ and using the same line of argument as in (3.13), one can derive similar approximations as EP-FACT or EP-1STEP. For example, $\tilde{p}^{\text{EP-FACT}}$

has the form

$$\tilde{p}^{\text{EP-FACT}}(\boldsymbol{x}_*) \propto \hat{q}(\boldsymbol{x}_*) \prod_j \int dx_j \hat{q}(x_i|\boldsymbol{x}_*)\epsilon_j(x_j).$$

One can check that the marginalization and the conditioning of $\hat{q}$ boils down to rank k updates, where k is the dimensionality of $\boldsymbol{x}_*$. For $k = 1$, the complexity $\tilde{p}^{\text{EP-FACT}}(x_*)$ roughly scales with the complexity of $\tilde{p}_i^{\text{EP-FACT}}(x_i)$.

3.4.5 Comparisons on toy models

In the following, we compare the performance of the marginal approximations on a few low-dimensional toy models; complex real-world models are considered in Section 3.6. For most of the models presented below, we use a prior p_0 with a symmetric covariance matrix $\boldsymbol{V} = v\big[(1-c)\boldsymbol{I} + c\mathbf{1}\mathbf{1}^T\big]$, where we vary the variance v and the correlation c. We have chosen the models below, because they are often used in practice, and they lead to sufficiently non-Gaussian posterior marginals.

Probit terms. The terms t_j are defined as $t_j(x_j) = \Phi(y_j x_j)$, where Φ is the Gaussian cumulative density function. This choice of terms is typically made in binary classification models, where $y_j \in \{-1, 1\}$. In order to obtain skewed marginals, in this example we set $y_j = 4$. The top and center panels in Figure 3.2 show the marginal corrections of the first component for a three-dimensional model with $(v, c) = (1, 0.25)$ and $(v, c) = (4, 0.9)$, respectively. The bars, in this and all other figures, correspond to a large number of Monte Carlo samples, either obtained through Gibbs or Metropolis sampling, and are supposed to represent the gold standard. The local correction EP-L yields sufficiently accurate approximations when the correlations are weak (top), but is clearly insufficient when they are strong (center). The corrections EP-1STEP and EP-FACT yield accurate estimates and are almost indistinguishable even for strong prior correlations. Only when we increase the number of dimensions (here from 3 to 32) and use strong prior correlations with moderate prior variances $(v, c) = (4, 0.95)$, we can see small differences (top-right). As we can see in Figure 3.2, EP-OPW performs slightly worse than EP-FACT and can indeed turn negative.

It is known that the Laplace method does not perform well on this model (e.g. Kuss and Rasmussen, 2005). The approximations it yields tend to be acceptable for weak correlations (top), with LA-CM and LA-FACT clearly outperforming LM and LM-L, but are far off when the correlations are stronger (center, bottom). These corrections suffer from essentially the same problems as the global Gaussian approximation based on Laplace's method: the mode and the inverse Hessian represent the mean and the covariance badly and fail to sufficiently improve it. It is interesting to see that LA-CM2 can be almost as accurate as LA-TK , while its computational complexity scales with LA-CM. The examples suggest that, at least in case of this model, LA-CM2 has the best accuracy/complexity tradeoff when compared to LA-CM and LA-TK .

Step-function terms. Expectation propagation can still be applied when it makes no sense to use the Laplace method. One such example is when the terms t_j are defined as $t_j(x_j) = \Theta(y_j x_j)$, where Θ is the step-function $\Theta(z) = \text{sign}(z)$ for $z \neq 0$ and $\Theta(0) = 1$. We chose $y_i = 1$. The plots on the left of Figure 3.3 show the marginals

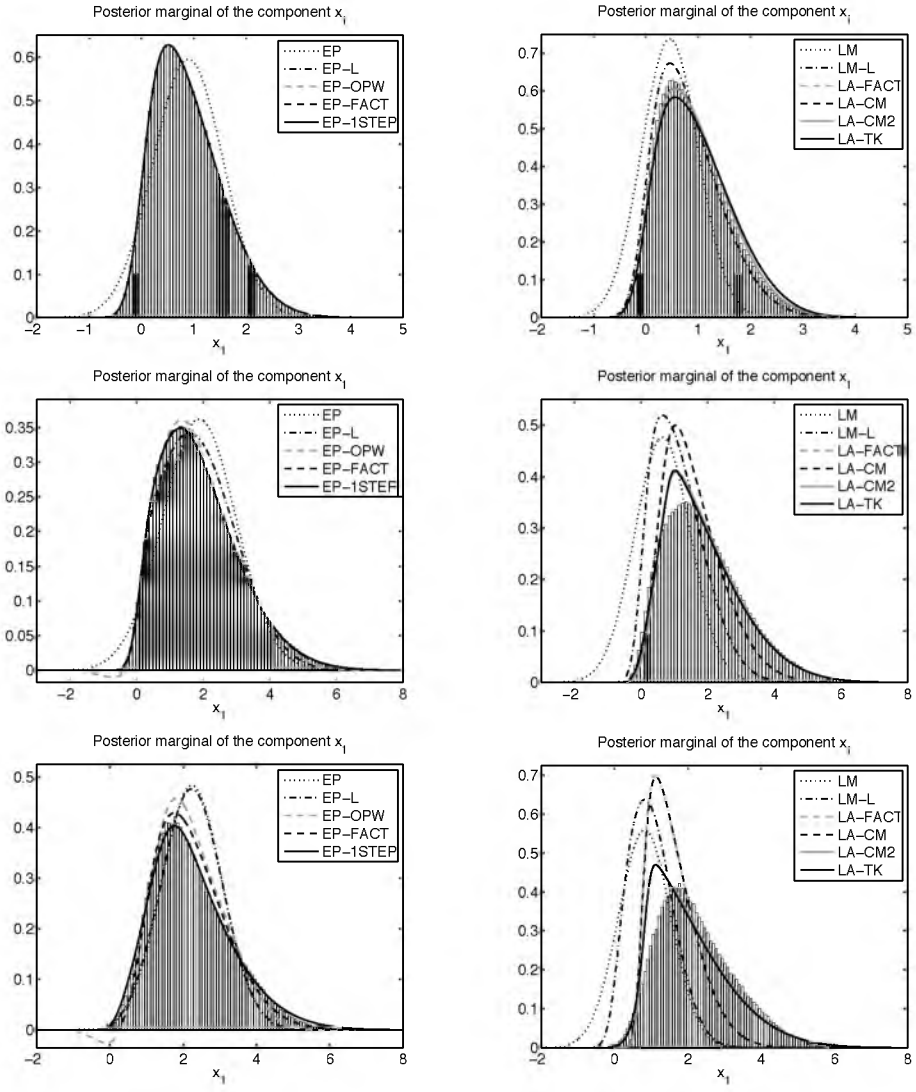

Figure 3.2: Various marginal corrections for a probit model with $t_i(x_i) = \Phi(4x_i)$ and identical variances and correlations in the prior p_0, using expectation propagation (left column) and Laplace approximations (right column). The panels show the corrections for a 3-dimensional model with prior variances and correlations $(v, c) = (1, 0.25)$ (top), $(v, c) = (4, 0.9)$ (center) and for a 32-dimensional model $(v, c) = (4, 0.95)$ (bottom). Note how, the accuracy of the approximations decreases as the correlation, the prior variance and the dimension of the problem increases.

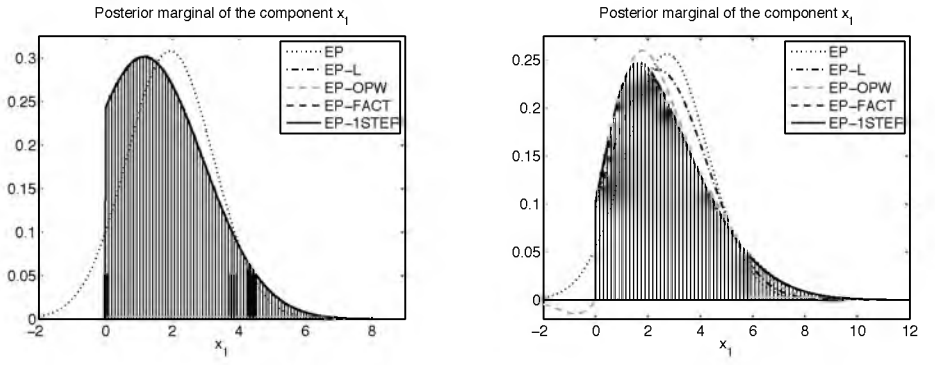

Figure 3.3: The posterior marginals of the first components of a 3-dimensional model with Heaviside terms with $(v, c) = (4, 0.5)$ (left) and $(v, c) = (9, 0.95)$ (right). The EP based approximations perform well even when the Laplace method is not applicable. The approximations have a similar behavior as in case of the probit model.

of the first component of a three dimensional model with $(v, c) = (4, 0.5)$ (left) and $(v, c) = (9, 0.95)$ (right). The performance of the approximations is similar to those of the previous model, except that in this case, we are dealing with non-continuous marginals.

Double-exponential terms. Another model where the Laplace method can fail, is the model where the terms t_j are defined as $t_j(x_j) = \lambda e^{-\lambda|y_j - x_j|}/2$, with $\lambda > 0$. The posterior marginals and their approximations for a three-dimensional model with $(v, c) = (9, 0.9)$ and $\lambda = 0.25$, $[y_1, y_2, y_3] = [-3, 0, 1]$ are shown on the panels of Figure 3.4. The marginals of the global EP approximation get the mass right, but not the shape. Local corrections already help a lot, while EP-OPW, EP-FACT and EP-1STEP are practically indistinguishable from the sampling results.

Linear regression with sparsifying prior. The double exponential distribution can also be used as a sparsifying prior in a linear regression setting. We choose a model with $n = 8$ variables and $m = 8$ observations—m being close to n led to the most interesting posterior marginals. The elements of the design matrix $\boldsymbol{U}$ are sampled according to the standard normal density and renormalized such that every column has unit length. The regression coefficients are chosen as $\boldsymbol{x} = [1, 1, 0, \ldots, 0]^T$ and the observations y_j are generated by $\boldsymbol{y} = \boldsymbol{U}\boldsymbol{x} + \boldsymbol{\epsilon}$, where ϵ_j is normal with variance $v = 0.01$. We take zero centered independent double exponential priors on the x_j coefficients. The panels of Figure 3.5 show a few posterior marginals of the regression coefficients x_j given the maximum a posteriori (MAP) hyper-parameters v and λ. The priors on the hyper-parameters are taken as independent and log-uniform. The approximations are accurate but in this case, the local approximations EP-L fail dramatically when the mass of the distribution is not close to zero.

A logistic regression model. We can try to use EP-FACT to approximate the marginal probabilities even when the terms t_i, $i \in \{1, \ldots, m\}$ depend on more than one variable. As an example, we define the terms as $t_i(\boldsymbol{x}) = \Phi(\boldsymbol{u}_i^T \boldsymbol{x})$. In this case, the factorization principle does not apply, but we can still use the line of argument in Section 3.4.2 and evaluate how EP-FACT performs. The panels of Figure 3.6 show a few marginals of a

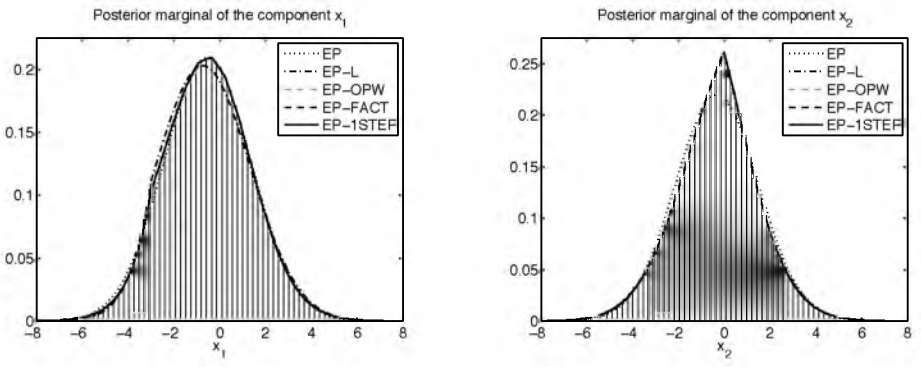

Figure 3.4: The posterior marginals of a three-dimensional model with $t(x_j) = \lambda e^{-\lambda|y_j - x_j|}/2$ ($\lambda = 0.25, [y_1, y_2, y_3] = [-3, 0, 1]$) and identical variances and correlations in p_0, corresponding to a prior variance and correlation $(v, c) = (9, 0.9)$.

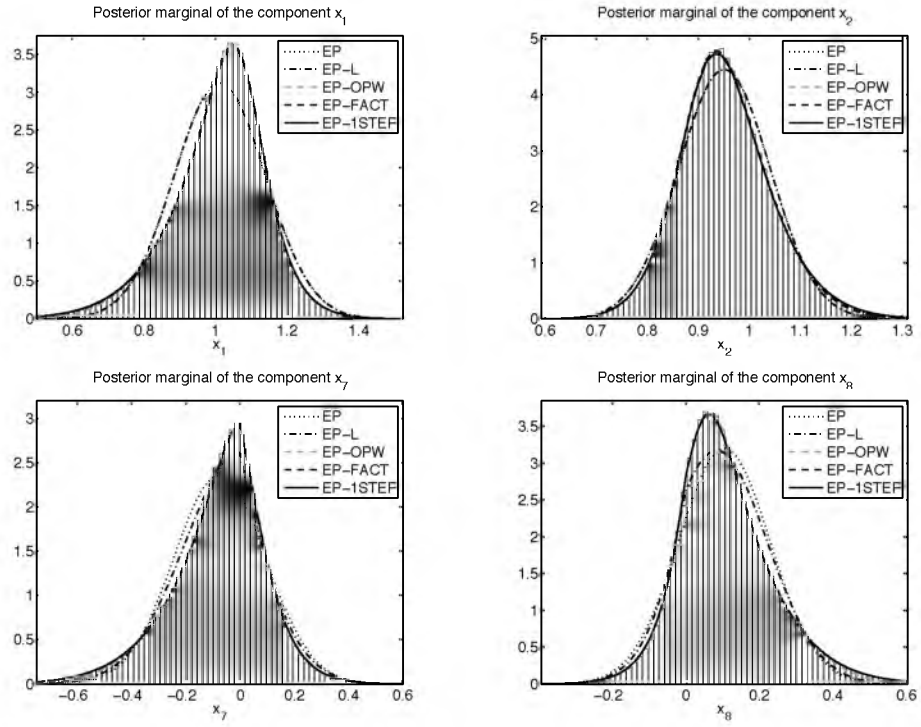

Figure 3.5: The posterior densities of a non-zero and a zero coefficient in a toy linear regression model with double exponential prior on the coefficients. It is interesting to compare the effects of the double exponential prior terms centered a zero on the quality of the local approximation EP-L. The effect is insignificant in the case the non-zero coefficient while in the case of the zero coefficient it has a strong effect, but the EP-L might still be quite inaccurate. We considered $n = 8$ coefficients the first two being 1 and the rest 0 and we generated $m = 8$ observables according to the model.

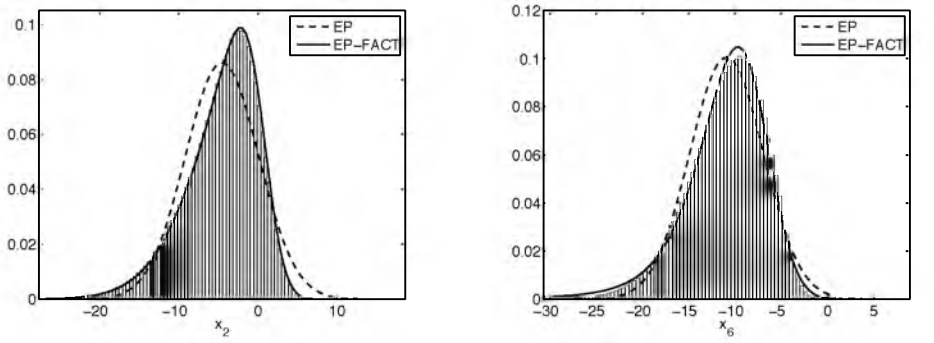

Figure 3.6: The posterior marginal approximation EP-FACT of the coefficients in a toy logistic regression model with Gaussian prior on the coefficients and moderate posterior correlations.. The panels show that even when the non-Gaussian terms depend on more than one variable and the posterior the approximation EP-FACT might still be accurate. We generated $n = 8$ coefficients and $m = 8$ observable variables.

model where we have chosen $w_i^j \sim \mathcal{N}(0, 10)$ and an independent Gaussian prior density $p_0(\boldsymbol{x}) = \prod_j N(x_j|0, v^{-1})$ with $v = 0.01$. We used $n = 8$ and $m = 8$. The correlations are moderate and in this case, EP-FACT seems to approximate well the marginals.

3.4.6 Computational complexities of the global approximations in sparse Gaussian models

In this section, we review the computational complexities of the Laplace method and expectation propagation when applied to sparse Gaussian models, i.e., models for which the n-dimensional precision matrix $\boldsymbol{Q}$ of the Gaussian prior is sparse. This is common in many practical applications in which the prior p_0 can be defined as a Gaussian Markov random field (e.g. van Gerven et al., 2009, 2010). We explore whether EP is indeed orders of magnitude slower, as suggested in Rue et al. (2009).

The computational complexity for both the (global) Laplace method and expectation propagation is dominated by several operations. 1) Computing the *Cholesky factor* $\tilde{\boldsymbol{L}}$ of a matrix $\tilde{\boldsymbol{Q}}$, e.g., corresponding to the posterior approximation $\tilde{p}^{\text{EP}}$ or $\tilde{p}^{\text{LM}}$, with the same sparsity structure as the prior precision matrix $\boldsymbol{Q}$. The computational complexity, denoted c_{chol}, scales typically with $nnzeros(\boldsymbol{Q})^2/n$, with $nnzeros(\boldsymbol{Q})$ being the number of non-zeros in the precision matrix $\boldsymbol{Q}$. 2) Computing the *diagonal elements of the inverse* of $\tilde{\boldsymbol{Q}}$. For sparse matrices, these can be computed efficiently by solving the Takahashi equations (Takahashi et al., 1973; Erisman and Tinney, 1975), which take the Cholesky factor $\tilde{\boldsymbol{L}}$ as input. A detailed description of solving the Takahasi equations can be found in Section A.2 of the Appendix. The computational complexity, denoted c_{taka}, scales with n^3 in the worst case, but typically scales with $nnzeros(\boldsymbol{L})^2/n$. In practice, we experienced that it is significantly more expensive than the Cholesky factorization, possibly due to the additional covariance values one has to compute during the process[2]. 3) Solving a

[2]We used the MATLAB implementation of the sparse Cholesky factorization and a C implementation for

triangular system of the form $\tilde{\boldsymbol{L}}\boldsymbol{a} = \boldsymbol{b}$, with corresponding computational complexity $c_{\text{tria}} \propto nnzeros(\boldsymbol{L})$.

The complexity of the latter two operations strongly depends on the number of non-zeros in the Cholesky factor, which should be kept to a minimum. There are various methods to achieve this by reordering the variables of the model. The approximate minimum degree reordering algorithm (Amestoy et al., 1996) seems to be the one with the best average performance (Ingram, 2006). Since the sparsity structure is fixed, the reordering algorithm has to be run only once, prior to running any other algorithm.

The Laplace method

To compute the global Gaussian approximation using the Laplace method, we first have to find the maximum a-posteriori solution. This can be done using, for example, the Newton method. Each Newton step requires one Cholesky factorization and solving two triangular systems. The off-diagonal elements of the posterior precision matrix $\tilde{\boldsymbol{Q}}$ are by construction equal to the off-diagonal elements of the prior precision matrix, so we only have to compute the n diagonal elements. To arrive at the lowest-order marginals $\tilde{p}_i^{\text{LM}}$ for all nodes i, we need the diagonal elements of the covariance matrix, the inverse of the precision matrix. These can be computed by solving the Takahashi equations, for which we can use the Cholesky factor computed in the last Newton step. Thus, computing the lowest order (Gaussian) marginals $\tilde{p}_i^{\text{LM}}$ for all variables $x_i,\ i = 1, \ldots, n$ by the Laplace method scales in total with $n_{\text{steps}}^{\text{Newton}} \times (c_{\text{chol}} + 2 \times c_{\text{tria}}) + c_{\text{taka}}$.

Expectation propagation

In order to update a term approximation $\tilde{t}_j\left(x_j\right)$, we compute $q^{\backslash j}\left(x_j\right)$ using the marginals $q\left(x_j\right)$ from the current global approximation $q\left(\boldsymbol{x}\right)$ and re-estimate the normalization constant and the first two moments of $t_j\left(x_j\right) q^{\backslash j}\left(x_j\right)$. In standard practice, the term approximations $\tilde{t}_j$ are updated sequentially and all marginal means and variances are recomputed using rank one updates after each term update. Instead, we adopt a parallel strategy, that is, we recompute marginal means and variances only after we have updated *all* term approximations $\tilde{t}_j$, with $j = 1, \ldots, n$.

A parallel EP step boils down to: 1) compute the Cholesky factorization of the current precision matrix, 2) solve two triangular systems to compute the current posterior mean and solve the Takahashi equations to compute the diagonal elements of the covariance matrix, and 3) if necessary, use univariate Gauss-Hermite numerical quadrature with n_{quad} nodes to compute the moments of $\epsilon_j(x_j)q(x_j)$ for all $j = 1, \ldots, n$. This adds up to a computational complexity that scales with $n_{\text{steps}}^{\text{EP}} \times (c_{\text{chol}} + 2 \times c_{\text{tria}} + c_{\text{taka}} + n \times n_{\text{quad}})$. After convergence, EP yields the lowest order Gaussian marginals $\tilde{p}_i^{\text{EP}}$ for all variables $x_i,\ i = 1, \ldots, n$.

Because of the parallel schedule, we can make use of exactly the same computational tricks as with the Laplace method (Cholesky, Takahashi). Since solving the Takahashi equations for large n dominates all other operations, the main difference between the Laplace method and EP is that for EP we have to solve these equations a number of times, namely the number of EP steps, yet for Laplace only once. Initializing the term

solving the Takahashi equations.

steps \ methods	LA-CM	LA-FACT	EP-1STEP	EP-FACT
$q(x_j\|x_i)$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$
$\tilde{\epsilon}(x_j; x_i)$	$n \times n_{\text{grid}}$	$n \times n_{\text{grid}}$	$n \times n_{\text{grid}} \times n_{\text{quad}}$	$n \times n_{\text{grid}} \times n_{\text{quad}}$
Norm. or det.-s	$c_{\text{chol}} \times n_{\text{grid}}$	$n \times n_{\text{grid}}$	$c_{\text{chol}} \times n_{\text{grid}}$	$n \times n_{\text{grid}}$

Table 3.1: Computational complexities of the steps for computing an improved marginal approximation for a particular node i using the various methods. The frames highlight the complexities that typically dominate the computation time. c_{tria}, c_{chol}, and c_{taka} refer to solving a sparse triangular system, a Cholesky factorization, and Takahashi equations, respectively. n_{grid} refers to the number of grid points and n_{quad} to the number of Gauss-Hermite quadrature nodes for x_i.

approximations in EP to the terms obtained by the Laplace method and then performing a few EP steps to obtain better estimates of the probability mass, makes EP just a (small) constant factor slower than Laplace. For efficient sequential updating of EP, we would need a fast one-rank Takahashi update (or something similar), which, to the best of our knowledge, does not exist yet.

It is interesting to realize that since for any $Q_{ij} \neq 0$ the Takahashi equations also provide $\left[\boldsymbol{Q}^{-1}\right]_{ij}$, we can run EP using the factors $t_{ij}(x_i, x_j) = t_i(x_i)^{1/n_i} t_j(x_j)^{1/n_j}$ where n_k is the number of neighbors of node k according to the adjacency matrix defined by the structure of $\boldsymbol{Q}$. This increases the amount of computation, but the approximation might be more accurate.

3.4.7 Computational complexities of marginal approximations

After running the global approximation to obtain the lowest order approximation, we are left with some Gaussian $q(\boldsymbol{x})$ with known precision matrix, a corresponding Cholesky factor and single-node marginals $q(x_i)$. We now consider the complexity of computing a corrected marginal through the various methods for a single node i, using n_{grid} grid points (see the summary in Table 3.1).

The local corrections $\tilde{p}_i^{\text{LM-L}}$ and $\tilde{p}_i^{\text{EP-L}}$ we get more or less for free. All other correction methods require the computation of the conditional densities $q(x_j|x_i)$. The conditional variance is independent of x_i, the conditional mean is a linear function of x_i. Computing $q(x_j|x_i)$ at all grid points for each j then amounts to solving two sparse triangular systems and $(n-1) \times n_{\text{grid}}$ evaluations. To arrive at the term approximations $\tilde{\epsilon}(x_j; x_i)$, we need to compute second order derivatives for the Laplace approximation and numerical quadratures for EP, which is about n_{quad} times more expensive. For LA-FACT, EP-OPW and EP-FACT, we then simply have to compute a product of n normalization terms. For LA-CM and EP-1STEP, we need to compute the determinant of an $(n-1)$-dimensional sparse matrix, which costs a Cholesky factorization. For LA-CM2 an additional c_{tria} has to be added for each x_i.

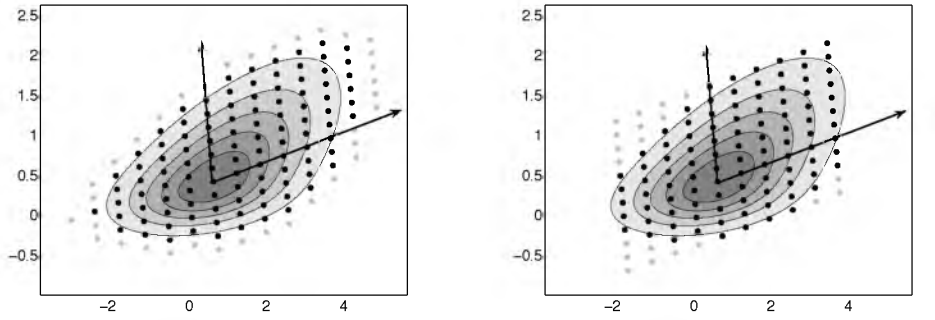

Figure 3.7: A comparison of the points selected by the thresholding breadth-first search procedure (left panel) and the method proposed by Rue et al. (2009) (right panel) when exploring in the eigenspace corresponding to the modal configuration. The black dots show the selected points while the gray ones stand for the ones that do not satisfy the thresholding condition. The principal axes on the figure are not perpendicular because of the different scaling of the axes. The number of evaluations in our method roughly grows proportional to the volume of a d-dimensional sphere, whereas the method of Rue et al. (2009) relates to the (larger) volume of a d-dimensional cube.

3.5 Inference of the hyper-parameters

Until now, we considered estimating single-node marginals conditioned upon the hyper-parameters. In this section, we consider the estimation of the posterior marginals that follow by integrating over the hyper-parameters. For this, we need the posterior density of the hyper-parameters given the observations, which is approximated by $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$, where $\tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})$ is the evidence approximation provided by the Laplace method or expectation propagation. For the moment we assume that the approximate posterior density of the hyper-parameters is unimodal.

We propose a slight modification of the method used by Rue et al. (2009). Their method explores the space of the hyper-parameters in the eigen-space corresponding to the modal configuration and can be described briefly as: (1) compute the modal configuration $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of $\log \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$, (2) starting from the mode $\boldsymbol{\mu}$, select a set of uniformly spaced nodes $\mathcal{X}_i$ along the scaled eigenvectors $\sqrt{\lambda_i}\boldsymbol{u}_i$—here $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$—by thresholding at both ends according to $\log \tilde{p}(\boldsymbol{\mu}|\boldsymbol{y}) - \log \tilde{p}\left(\boldsymbol{\mu} + k_i \Delta \sqrt{\lambda_i}\boldsymbol{u}_i|\boldsymbol{y}\right) < \delta, k_i \in \mathbb{Z}$, and finally (3) use all hyper-parameters corresponding to the nodes of the product grid $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$, $d = \dim(\boldsymbol{\theta})$ and satisfying the latter thresholding condition, to perform numerical quadrature using the rectangle rule.

Since the computational bottleneck of the procedure is the evaluation of the approximate evidence, we propose to improve this method by selecting the nodes—step (2) from above—in a different way: we keep the thresholding condition but we do a breadth–first search with regard to $(k_1, \ldots, k_d)$ on the grid graph $\mathbb{Z}^d$. We start from the origin and the hyper-parameter values that do not satisfy the thresholding condition are not included in the set of nodes whose neighbors we search. This simple modification proves to be very economical, since when exploring the volume around the mode, only the hyper-parameters that form the boundary surface are explored, but not selected. Thus, the pro-

portion of useless computational time is the ratio of surface to volume. Although the boundary nodes do not satisfy the thresholding conditions, we can still use them in the numerical procedure. The number of grid points to be evaluated grows exponentially, as it does for the method in Rue et al. (2009). The difference is that in our method it roughly grows proportional to the volume of a d-dimensional sphere, whereas in the case of the method in Rue et al. (2009) it relates to the (larger) volume of a d-dimensional cube.. When the posterior density is not unimodal then we suggest to use a uniform d-dimensional uniformly spaced grid, that is, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and choose a well suited $\boldsymbol{\mu}$ and threshold δ which allows the exploration of the most significant modes. Figure 3.7 illustrates the breadth-first search method on two–dimensional example compared to the method proposed to Rue et al. (2009). Once the hyper-parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ are selected, the integration of the corrected approximate marginals over the hype-parameter's approximate posterior density can be written as

$$\tilde{p}\left(x_i|\boldsymbol{y}\right) = \frac{\sum_{j=1}^{m} \tilde{p}\left(x_i|\boldsymbol{y}, \boldsymbol{\theta}_j\right) \tilde{p}\left(\boldsymbol{\theta}_j|\boldsymbol{y}\right)}{\sum_{j=1}^{m} \tilde{p}\left(\boldsymbol{\theta}_j|\boldsymbol{y}\right)}, \tag{3.19}$$

implying that the proposed procedure is similar to a reasonably efficient sampling procedure.

3.6 Examples

As real-world examples, we chose two models: one from Zoeter and Heskes (2005b) and one from Rue et al. (2009). Our aim was to show that the EP based correction methods can be as accurate as the Laplace approximation based ones and given that we have a sparse Gaussian prior, EP can be considered as an alternative to the Laplace method even when the number of variables is of the order of tens of thousands.

3.6.1 A stochastic volatility model

As a first example for a sparse Gaussian model, we implemented the stochastic volatility model presented in Zoeter and Heskes (2005b) where the they applied a sequential (global) EP to approximate the posterior density. The same model was used by Rue et al. (2009) to show that the global Laplace approximation is by magnitudes faster in sparse models than a sequential EP. They also showed that their marginal approximations work well on this model.

The data set consists of 945 samples of the daily difference of the pound-dollar exchange rate from October 1^{st}, 1981, to June 28^{th}, 1995. The observations y_t given the latent variables η_t are taken to be distributed independently according to $p\left(y_t|\eta_t\right) = N\left(y_t|0, e^{\eta_t}\right)$. The quantity η_t governing the volatility is a linear predictor defined to be the sum $\eta_t = f_t + \mu$ of a first-order auto-regressive Gaussian process $p\left(f_t|f_{t-1}, \phi, \tau\right) = N\left(f_t|\phi f_{t-1}, 1/\tau\right)$, with $|\phi| < 1$, and an additional Gaussian bias term wit a prior $\mu \sim N\left(\mu|0, 1\right)$. Thus the prior on $(f_1, \ldots, f_T, \mu)$ is a sparse latent Gaussian field. The prior on the hyper-parameter τ is taken to be $p\left(\tau\right) = \Gamma\left(\tau|1, 10\right)$ and a Gaussian prior $\mathcal{N}\left(0, 3\right)$ is taken over $\phi' = \log\left((1+\phi)/(1-\phi)\right)$.

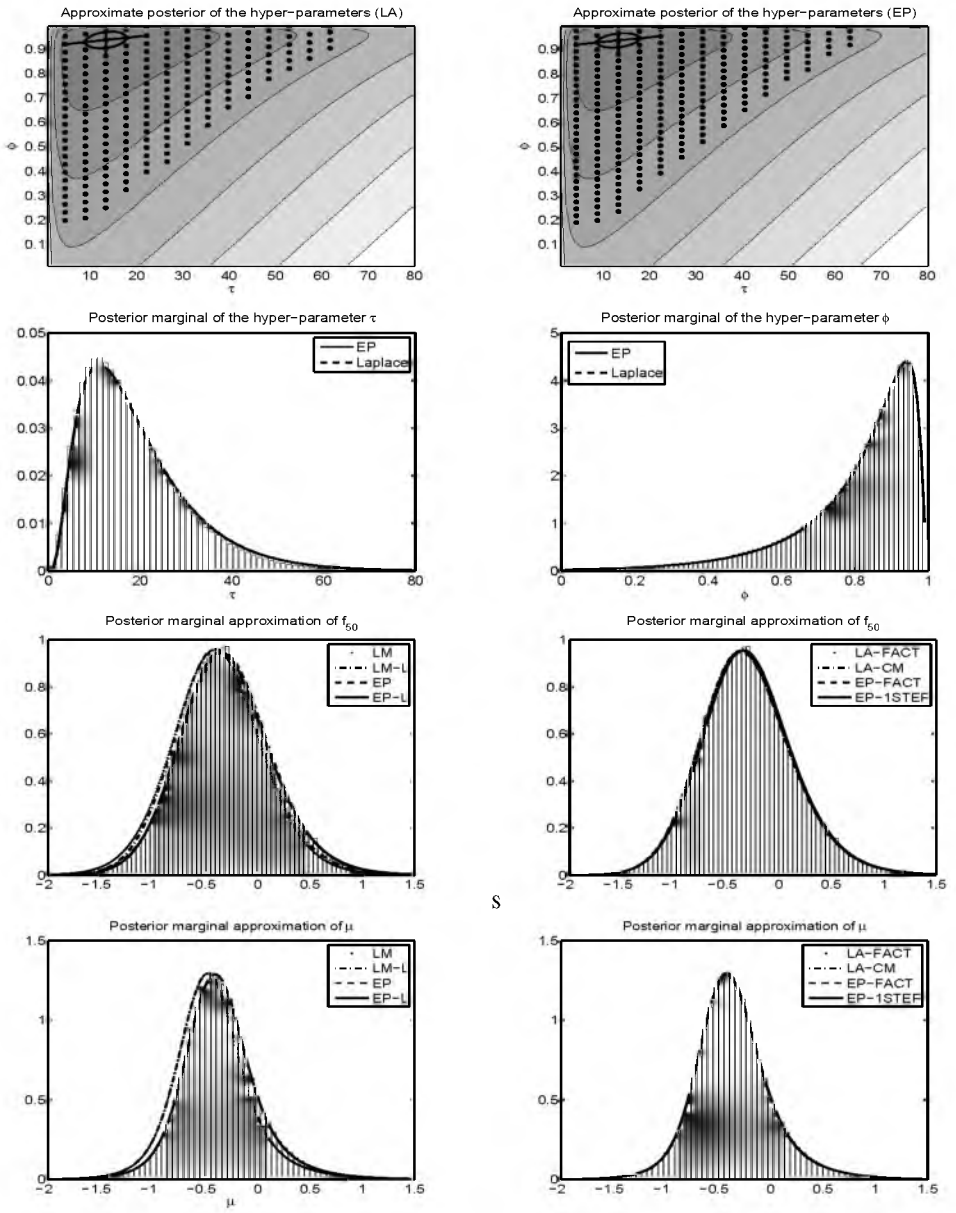

Figure 3.8: Plots of the posterior densities in the stochastic volatility model in Section 3.6.1. Figure panels show the logarithm of the approximate posterior density of the hyper-parameters using EP (top-right) and the Laplace method (top-left), their marginals (second row) and the posterior marginal approximations of f_{50} and μ (bottom rows) when integrated over the corresponding approximations of the hyper-parameters' posterior density. Dots show the hyper-parameters used for numerical integration; ellipses visualize the Hessian at the approximate posterior density's mode. The rest of the panels show the posterior density approximations of f_{50} and μ.

The joint density of the stochastic volatility model is

$$p(\boldsymbol{y}, \boldsymbol{f}, \mu, \tau, \phi) = \prod_{t=1}^{T} N\left(y_t|0, e^{f_t+\mu}\right) N\left(f_1|0, 1\right) \prod_{t=2}^{T} N\left(f_t|\phi f_{t-1}, 1/\tau\right) \tag{3.20}$$
$$\times N\left(\mu|0, 1\right) \Gamma\left(\tau|1, 10\right) N\left(\log\left(\frac{1+\phi}{1-\phi}\right)|0, 3\right)\left(\frac{2}{1-\phi^2}\right),$$

where $\Gamma\left(\cdot|k, \theta\right)$ denotes the Gamma density with mean value $k\theta$. Rue et al. (2009) propose to use the the first 50 observations, as this is the regime where the posterior marginals have the most interesting behavior. For comparison, we used the same number of observations.

The results are shown in Figure 3.8. The Laplace and EP approximation of the evidence are nearly indistinguishable (top-row), as are the posterior marginals of the hyperparameters (second row). Here EP is around a factor 5 slower than Laplace. The posterior marginals of f_{50} and μ obtained using the more involved methods (bottom rows) are practically indistinguishable from each other and the gold (sampling) standard. This is not the case for the cheaper variants LM, EP, and LM-L, but *is* the case for EP-L (third row): apparently to obtain excellent posterior marginals on this model, there is no need for (computationally expensive) corrections, but it suffices to compute a single global EP approximation per hyper-parameter setting and correct this for the (non-Gaussian) local term.

3.6.2 A log-Gaussian Cox process model

As a large sized example, we implemented the Laplace approximation and expectation propagation for the log-Gaussian Cox process model applied to the tropical rainforest biodiversity data as presented in Rue et al. (2009). The observational data used in Rue et al. (2009) is the number of trees y_{ij} form a certain species in a small rectangular rainforest area indexed by $i = 1, \ldots, 21$ and $j = 1, \ldots, 11$ with mean altitude a_{ij} and gradient g_{ij}. The data is modeled by a discretized Poisson point process in two dimensions and the log of the mean parameter η_{ij} is defined as a Gaussian field. This means that the observations y_{ij} are taken to be Poisson distributed with mean $w_{ij}e^{\eta_{ij}}$, where the parameters w_{ij} are proportional to the size of the area where y_{ij} is measured. Since Rue et al. (2009) consider rectangular areas of the equal size, in their model w_{ij} is constant. The latent Gaussian field η_{ij} modeling the log of the mean is defined as

$$\eta_{ij} = \beta_a a_{ij} + \beta_g g_{ij} + \beta_0 + f_{ij}^{(s)} + f_{ij}^{(u)}$$

where a_{ij} and g_{ij} are scalar quantities specifying altitude and gradient data, β_a and β_g are the corresponding linear coefficients and β_0 is a bias parameter. The latent fields $\boldsymbol{f}^{(s)}$ and $\boldsymbol{f}^{(u)}$ are defined as follows: $\boldsymbol{f}^{(s)}$ is a second-order polynomial intrinsic Gaussian Markov random field with precision parameter e^{v_s} constructed to mimic a thin plate spline on a uniform two dimensional grid, while $\boldsymbol{f}^{(u)}$ is an independent field with $f_{ij}^{(u)} \sim \mathcal{N}(0, e^{-v_\eta})$ included to model the noise. Independent wide priors $\mathcal{N}(0, v_\beta^{-1})$ are taken on β_a, β_g and β_0, with $v_\beta^{-1} = 10^3$. We worked with the data set used in the INLA software package

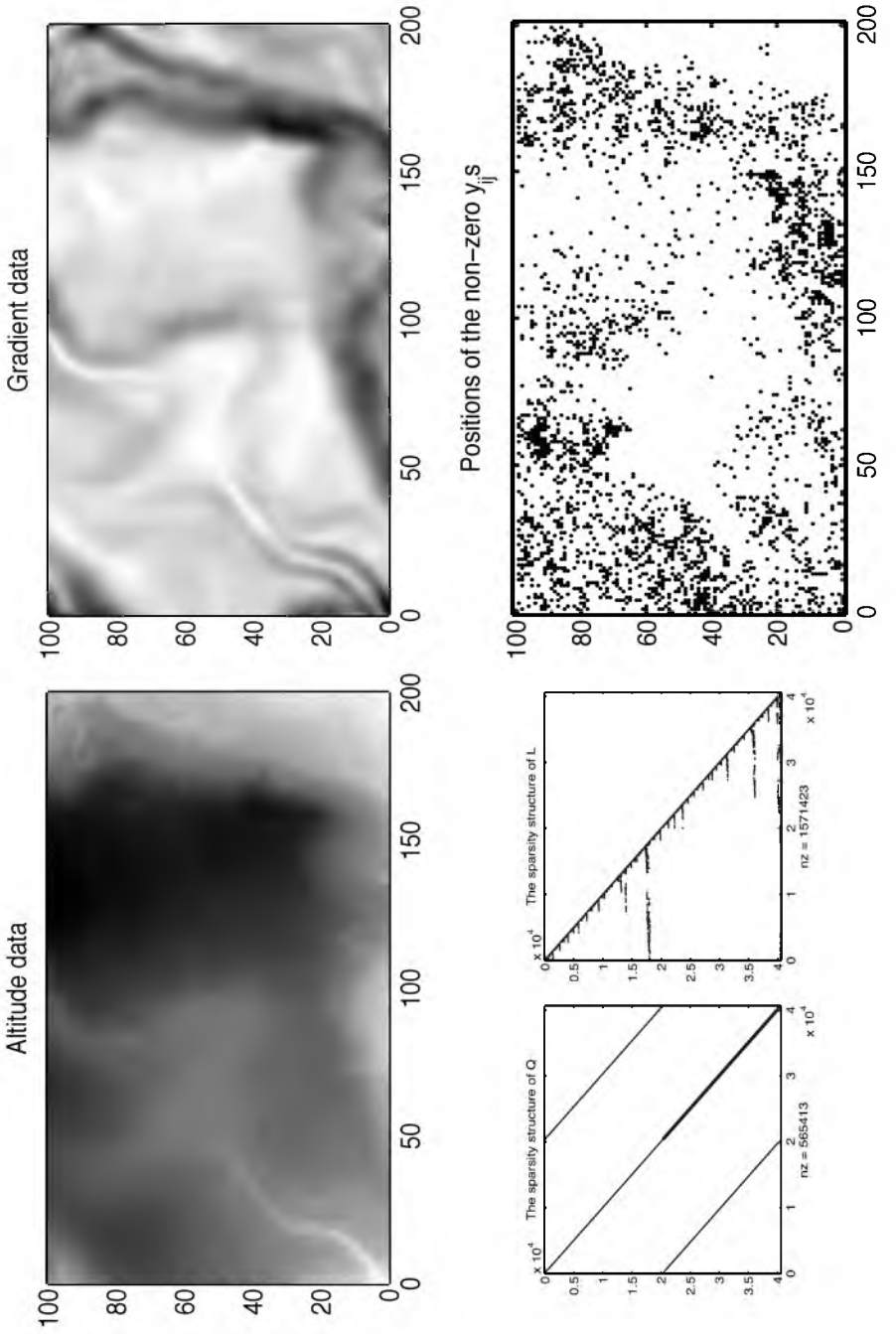

Figure 3.9: The panels show the altitude a_{ij}, gradient g_{ij} and the non-zero observation y_{ij} data for the log-Gaussian Cox process model in Section 3.6.2 together with the sparsity structure of $\boldsymbol{Q}$ and the Cholesky factor $\boldsymbol{L}$ of its approximate minimum degree reordering

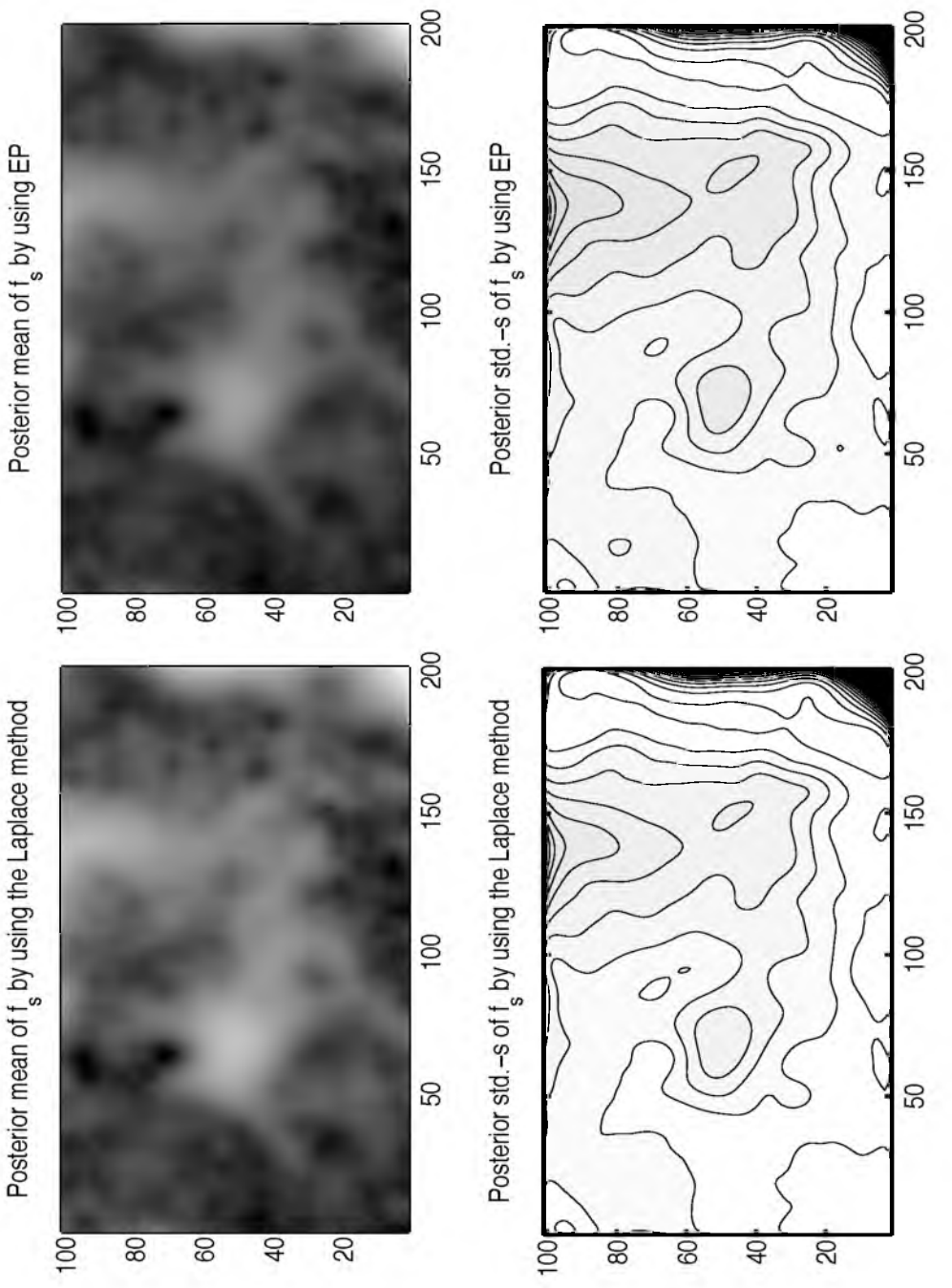

Figure 3.10: The posterior approximations of the evidence (top) and β_a and β_g (bottom). The Laplace method results in similar evidence estimates as EP (the level curves on the top panels show identical levels). The marginal approximations show marginals for the approximate MAP hyper-parameters..

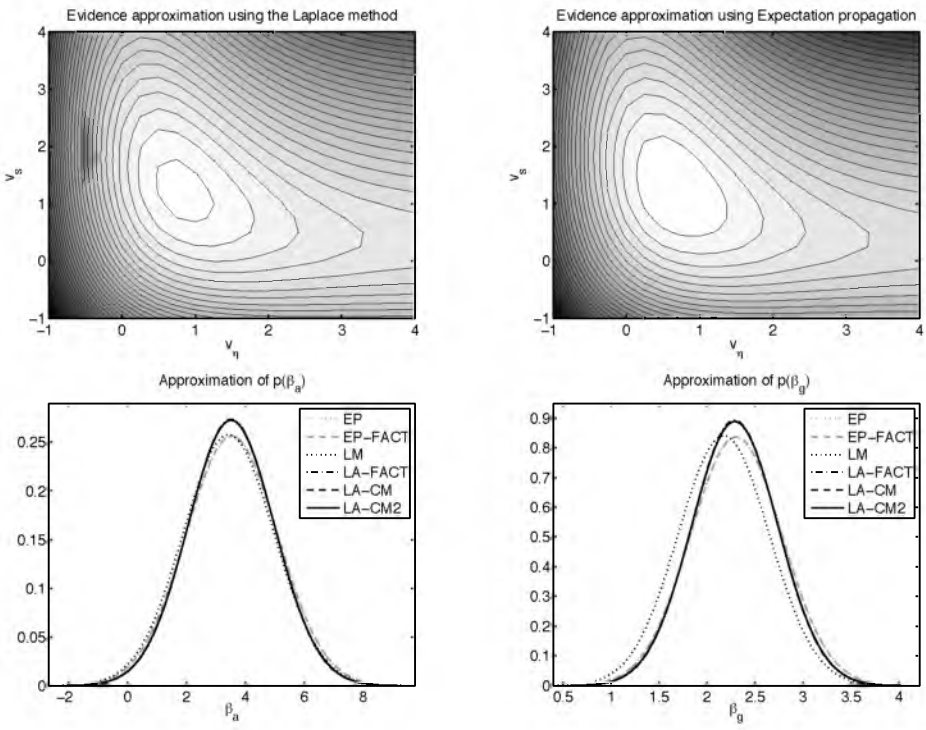

Figure 3.11: The posterior approximations of the evidence (top) and β_a and β_g (bottom). The Laplace method results in similar evidence estimates as EP (the level curves on the top panels show identical levels). The marginal approximations show marginals for the approximate MAP hyper-parameters.

(Martino and Rue, 2009). The data set contains the corresponding a_{ij}, g_{ij}, w_{ij} and y_{ij} for a grid size of 101×201. We also used the same modeling approach, that is, we have taken $(\boldsymbol{\eta}^T, \boldsymbol{f}^{(s)^T}, \beta_a, \beta_g, \beta_0)^T$ as latent variable, thus having an inference problem of dimension 40605. The joint density of the log-Gaussian Cox process model is

$$
\begin{aligned}
p(\boldsymbol{y},\boldsymbol{\eta}, \boldsymbol{f}^{(s)}, \beta_a, \beta_g, \beta_0 | v_\eta, v_s, \boldsymbol{a}, \boldsymbol{g}, \boldsymbol{w}) = \qquad (3.21)\\
= \prod_{ij} \text{Poisson}\left(y_{ij} | w_{ij} e^{\eta_{ij}}\right) N\left(\eta_{ij} | f_{ij}^{(s)} + a_{ij}\beta_a + g_{ij}\beta_g + \beta_0, v_\eta^{-1}\right)\\
\times \left(\frac{v_s}{2\pi}\right)^{N/2} |\boldsymbol{S}|_*^{1/2} \exp\left\{-\frac{1}{2} v_s \boldsymbol{f}^{(s)^T} \boldsymbol{S} \boldsymbol{f}^{(s)}\right\} N\left(\beta_a, \beta_g, \beta_0 | \boldsymbol{0}, 10^3 \boldsymbol{I}\right),
\end{aligned}
$$

where $|\boldsymbol{S}|_*$ is the generalized determinant—an irrelevant constant—of the structure matrix $\boldsymbol{S}$ consisting of the finite difference coefficients of a second order improper polynomial Gaussian Markov random field on a uniform two dimensional grid—with the corresponding boundary conditions (Rue and Held, 2005). We used uninformative priors for v_η and v_s. The bottom-right panels of Figure 3.9 show the sparsity structure of the precision matrix $\boldsymbol{Q}$ corresponding to the Gaussian random vector $(\boldsymbol{\eta}^T, \boldsymbol{f}^{(s)^T}, \beta_a, \beta_g, \beta_0)^T$ and the sparsity structure of its Cholesky factor $\boldsymbol{L}$ when $\boldsymbol{Q}$ is reordered with the AMD algorithm.

Expectation propagation was initialized using the term approximations corresponding to the Laplace method. Figure 3.9 shows the data we used and Figure 3.10 shows the mean values and standard deviations of $\boldsymbol{f}_s$ when using the Laplace method and EP with the hyper-parameter fixed to their corresponding approximate a posteriori (MAP) value.

The top panels of Figure 3.11 show the evidence approximations while the bottom panels show the marginal approximations for the corresponding MAP hyper-parameters. For β_a, there is a slight difference in variance between the Laplace approximation and the EP based methods, while for β_g,s besides a similar effect, the approximation methods also improve on the mean of LM. It seems that EP is a sufficiently good approximation and EP-FACT does not really improve on it.

3.6.3 A ranking model

To show that we can implement linear constraints with EP and that the factorization principle might work even in is cases when the non-Gaussian terms depend on more than one variable, we use a ranking model for rating players in sports competitions. The model is a simplified version of the models presented in Dangauthier et al. (2008) and Birlutiu and Heskes (2007) and we only consider it as an example to support the above mentioned claims. We assume that a player j is characterized by his/her strength which at time t is $x_t^{(j)}$. The prior on the evolution of the players' strength $\boldsymbol{x}_t = (x_t^{(1)}, \ldots, x_t^{(n)})$ is taken to be a factorizing $AR(1)$ model. Each game between two players is represented by the triple (i, j, t) and the collection of these triples is denoted by G. We assume that the outcomes of the games are a binary variables $y_{i,j,t} \in \{-1, 1\}$, the games are conditionally independent given the players strengths and the probability of player i winning the game against player j at time t is $\Phi(x_t^{(i)} - x_t^{(j)})$, where Φ is the standard normal cumulative density function. To implement the linear constraints, we constrain the players' strength

to sum to zero at any given time t. They are purely artificial and are only considered for illustration purposes.

The joint posterior density of the players' strength is given by

$$p(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^T | \boldsymbol{y}, v_1, v, a) \propto \prod_{t=1}^{T} \delta_0(\mathbf{1}^T \boldsymbol{x}_t) \prod_{(i,j,t)\in G} \Phi(y_{i,j,t}(x_t^{(i)} - x_t^{(j)})) \times \prod_{j=1}^{n} N(x^{(j)}|0, v_1) \prod_{t=1}^{T-1} N(x_{t+1}^{(j)}|a x_t^{(j)}, v).$$

We approximate this density with a Gaussian density using EP and we use the factorized corrections EP-FACT, to improve on the Gaussian marginals. The prior on the players strengths is a sparse Gaussian Markov random field, thus we can apply the methods presented in Section 3.4.6.

We have chosen a dataset consisting of four[3] tennis players and their ATP tournament games played against each other form 1995 to 2003. There was a total of 45 games played during these years. We run the model with a fixed set of parameters $v_1 = 1$, $a = 1$ and $v = 9$. The left panel in Figure 3.12 shows the evolution of the players' mean strengths and the corresponding standard deviations for the best player. Note that the players' mean strengths average to zero at all times. The right panel shows that the factorized approximations EP-FACT, can indeed improve on the Gaussian marginal approximations computed by EP even in models where non-Gaussian terms depend on more than one variable. This might be due to the relatively sparse interaction between the variables $x_t^{(j)}$, $t = 1, \ldots, T$, $j = 1, \ldots, n$.

3.7 Discussion

We introduced several methods to improve on the marginal approximations obtained by marginalizing the global approximations. The approximation denoted by EP-FACT seems to be, in most cases, both accurate and fast. An improvement in accuracy can be achieved with some additional computational cost by using EP-1STEP. We showed that by using a parallel EP scheduling the computational complexity of EP in sparse Gaussian model can scale with the computational complexity of the Laplace method.

There are many options for further improvement, in particular with respect to efficiency. The ideas behind the simplified Laplace approximation of Rue et al. (2009), which aims to prevent the expensive computation of a determinant for each x_i, are applicable to expectation propagation. However, if the computation of the determinant in EP-1STEP dominates the computation time, the factorized approximation EP-FACT may be a faster but less accurate alternative.

One of the main problems of expectation propagation is that it is not guaranteed to converge and may run into numerical problems. EP converged fine on the problems considered in this chapter, but even when it does not, it can still be useful to start from the Laplace solution and perform a few EP steps to get a better grip on the probability mass instead of relying on the mode and the curvature.

[3]We have chosen A. Agassi, Y. Kafelnikov, C. Moya and T Henman.

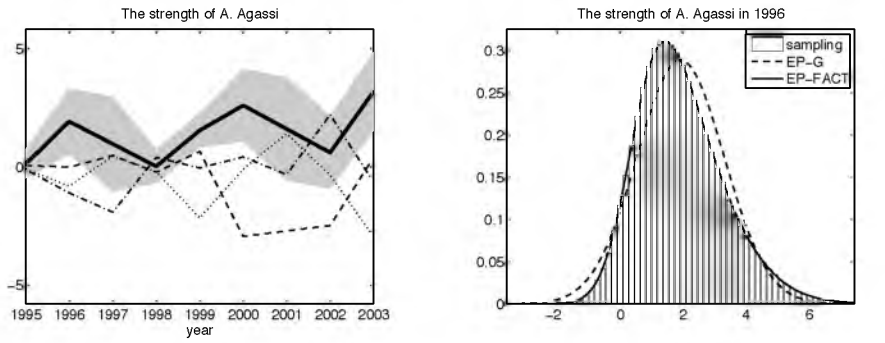

Figure 3.12: The left panel shows the mean strengths of the players A. Agassi (cont.), Y. Kafelnikov (dashed), C. Moya (dashed-dotted), and T. Henman (dotted) with the standard deviations of A. Agassi's strength based on the ranking model presented in Section 3.6.3. The dataset consists of the games played by these players against each other in the years 1995–2003. We implemented linear constraints such that the players strength sum to zero in every year. The left panel shows that this indeed holds for the means. The right panel shows A. Agassi's strength distribution in 1996 which is a non-Gaussian density and can be well approximated using EP-FACT.

For models with weak correlations and smooth nonlinearities, any approximation method gives decent results. It is possible to come up with cases (strong correlations, hard nonlinearities), where any deterministic approximation method fails. Most interesting problems are somewhere in between, and for those we can hardly tell how advanced and computationally intensive an approximation method we need. The heuristic suggested in Rue et al. (2009), to systematically increase the complexity and stop when no further changes can be obtained, appears to be risky. In particular when going from the factorized to the non-factorized approximations, it is often hard to see changes, but still both approximations can be far off. It would be interesting to obtain a better theoretical understanding of the (asymptotic) approximation errors implied by the different approaches.

3.8 A summary of the marginal approximation methods

An explanatory list of the approximation methods in Figure 3.13.

- EP-L. EP local. The approximation $\tilde{p}^{\text{EP-L}}(x_i) \propto \epsilon_i(x_i)q(x_i)$ is obtained from $c_{x_i}(x_) \approx 1$, where $\epsilon_i(x_i) = t_i(x_i)/\tilde{t}_i(x_i)$ and $q(\boldsymbol{x})$ are computed by EP (see Section 3.3).

- LM-L. Lapace method local. EP local. The approximation $\tilde{p}^{\text{EP-L}}(x_i) \propto \epsilon_i(x_i)q(x_i)$ is obtained from $c_{x_i}(x_) \approx 1$, where $\epsilon_i(x_i) = t_i(x_i)/\tilde{t}_i(x_i)$ and $q(\boldsymbol{x})$ are computed by the Laplace method (see Section 3.3). In this case $\log \epsilon_i(x_i) = R_2[\log t_i](x_i)$.

- EP-FULL. The full EP approximation of the marginal. This approximation is computed by using EP to approximate $c_i(x_i)$ (see Section 3.3.2).

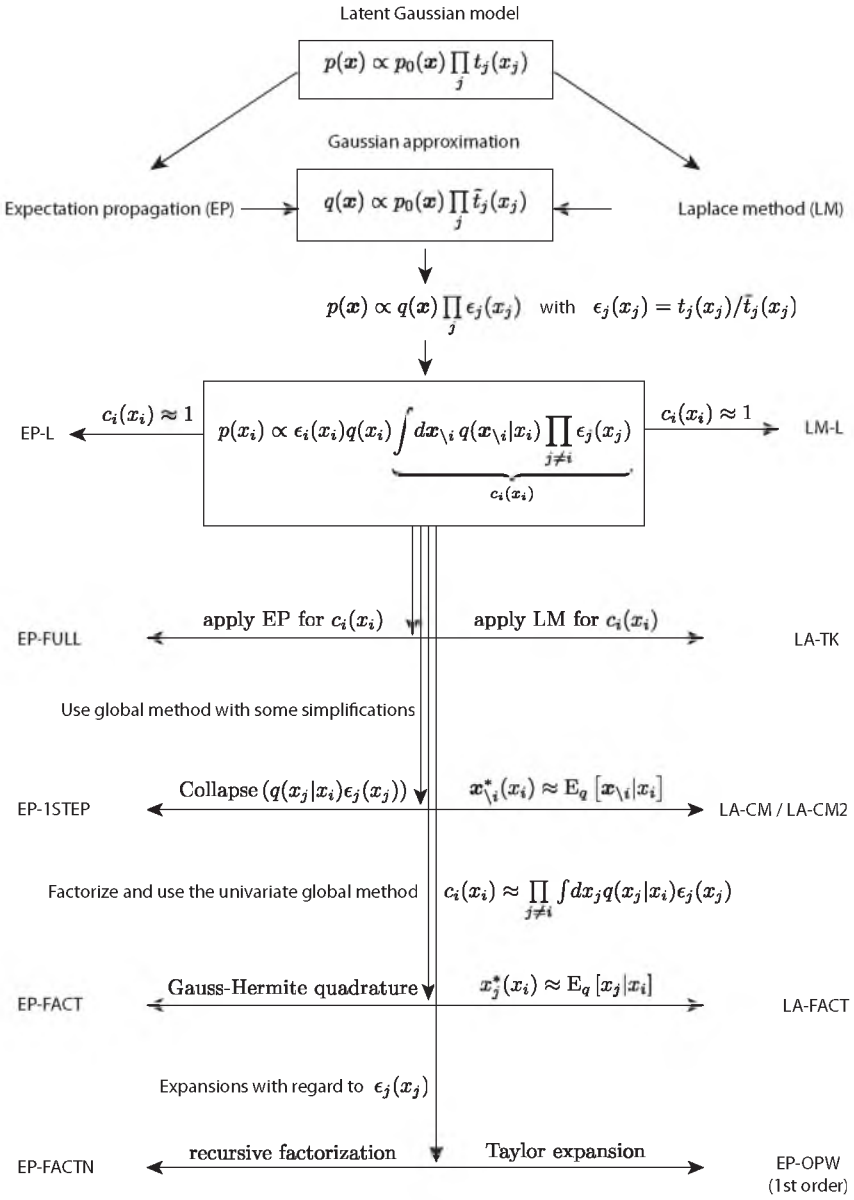

Figure 3.13: A schematic view of the approximation methods introduced or referred to in this paper. For details see Section 3.8.

- LA-TK . The Laplace approximation of Tierney and Kadane (1986). The approximation $\tilde{p}^{\text{LA-TK}}(x_i)$ is computed by using the Laplace method to approximate $c_i(x_i)$ (see Section 3.3.1).

- EP-1STEP. The one step EP approximation. The approximation $\tilde{p}^{\text{EP-1STEP}}(x_i)$ is computed by defining $\tilde{\epsilon}_j(x_j; x_i) \equiv \text{Collapse}(q(x_j|x_i)\epsilon_j(x_j))/q(x_j|x_i)$ and using the approximation $c_i(x_i) \approx \int d\boldsymbol{x}_{\setminus i}\, q(\boldsymbol{x}_{\setminus i}|x_i) \prod_{j\neq i} \tilde{\epsilon}_j(x_j; x_i)$ (see Section 3.4.1). This corresponds to the first EP step for computing $c_i(x_i)$ with the initialization $\tilde{\epsilon}_j(x_j; x_i) = 1$.

- LA-CM. The Laplace approximation with the conditional mode approximated by the conditional mean. The approximation $\tilde{p}^{\text{LA-CM}}(x_i)$ is computed as proposed in Rue et al. (2009), that is, by using the approximation $\boldsymbol{x}^*_{\setminus i}(x_i) \approx \mathrm{E}_q\big[\boldsymbol{x}_{\setminus i}|x_i\big]$ where $q(\boldsymbol{x})$ is given by the Laplace method (see Section 3.4.1).

- LA-CM2. The similar approximation as LA-CM, but with an additional term added to account for $\boldsymbol{x}^*_{\setminus i}(x_i) \approx \mathrm{E}_q\big[\boldsymbol{x}_{\setminus i}|x_i\big]$ (see Section 3.4.1).

- EP-FACT. The factorized EP approximation. The approximation $\tilde{p}^{\text{EP-FACT}}(x_i)$ is computed using the approximation $c_i(x_i) \approx \prod_{j\neq i} \int dx_j\, q(x_j|x_i)\epsilon_j(x_j)$, where the univariate integrals are computed numerically or analytically, if it is the case. For further details see Section 3.4.2.

- LA-FACT. A similar approximation as EP-FACT, but here, the univariate integrals are computed with the Laplace method and using the approximation $x^*_j(x_i) \approx \mathrm{E}_q[x_j|x_i]$, with $q(\boldsymbol{x})$ being the global approximation resulting from the Laplace method. For further details see Section 3.4.2.

- EP-OPW. The Taylor expansion of Opper et al. (2009). The approximation $\tilde{p}^{\text{EP-OPW}}(x_i)$ is computed by expanding $p(\boldsymbol{x}) \propto p_0(\boldsymbol{x}) \prod \epsilon_j(x_j)$ in first order with reagard to $\epsilon_j(x_j) - 1$ for all $j = 1, \ldots, n$ and integrating with regard to $\boldsymbol{x}_{\setminus i}$. When expanding only for $j \neq i$ this approximation is equal in first order to $\tilde{p}^{\text{EP-FACT}}(x_i)$ (see Section 3.4.3).

- EP-FACTN. Higher order approximations obtained by using the factorization recursively. For further details see Section 3.4.2.

Chapter 4

A multivariate sparsity inducing scale mixture prior

Summary

In this chapter, we introduce a multivariate sparsity inducing prior distribution that can be viewed as a multivariate generalization of the double exponential distribution. The distribution is constructed in a hierarchical way as a scale mixture distribution with the help of a multivariate exponential distribution. This approach leaves the variables uncorrelated, but it introduces correlations between their magnitudes. When applied in a linear regression and logistic regression setting, the symmetry properties of the distribution lead to posterior densities with block diagonal correlation structures. Our experiments on real-world MEG and fMRI data show that when used as a prior, the scale mixture distribution we introduce can take into account spatial and spatio-temporal smoothness properties and leads to meaningful smooth importance maps and fast approximate inference algorithms. The material presented in this chapter is based on van Gerven et al. (2009)[1] and van Gerven et al. (2010)[2].

4.1 Introduction

In many real-world models *interpretability* plays a similarly important role as *prediction accuracy*. Given a set of observations, it is often assumed that only a certain subset of the variables is responsible or relevant for producing the outcome, that is, there is a small or moderate sized subset of model parameters that exhibit the strongest effect. In this chapter, we study models with linear dependencies. In these models, the relevance can be measured by the magnitude of the regression coefficients: magnitudes (very) close to zero indicate irrelevance. These models are typically under-determined, that is, there is a large set of parameter settings that explain the data equally well. Depending on the

[1] M. van Gerven, B. Cseke, R. Oostenveld, and T. Heskes, *Bayesian source localization with the multivariate Laplace prior*, NIPS–2009, pages 1901–1909.

[2] M. van Gerven, B. Cseke, F. de Lange, and T. Heskes, *Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior*, Neuroimage, 50(1):150161, 2010.

problem we are dealing with, there are various techniques (e.g. Tibshirani, 1996) to select the best from these sets of parameters, a general paradigm being the preference for sparse parameter sets, that is, parameter sets where many regression coefficients are zero.

A way to implement this paradigm is to use maximum likelihood methods with penalties or constraints on the model parameters (e.g. Tibshirani, 1996; Williams, 1995; Cawley et al., 2007). The alternative way is the Bayesian framework (e.g. George and McCullogh, 1993; Seeger, 2008; Carvalho et al., 2009; Li and Lin, 2010), where after specifying a prior that expresses the preference for sparse parameter vectors one can compute the posterior probability density of the model parameters and make decisions by taking into account the (approximate) posterior uncertainties in the parameter values. Both methods have their advantages and drawbacks: the probabilistic setting allows us to assess the uncertainty in the estimates, but it does not lead to point estimates with (exactly) zero coefficients, while the regularized maximum likelihood framework leads to sparse point estimates and has mild constraints on defining the regularizer, but it gives no probabilistic interpretation of the results.

Most of the above-cited papers consider factorizing priors, or regularizers that are sums of regularizers on the individual regression coefficients. However, in many practical applications the model parameters (the regression coefficients) have an a priori (spatial) pattern, because they express effects that are (spatially) correlated (e.g. Penny et al., 2005). For example, in fMRI experiments it is reasonable to assume that the activation level of neighboring brain areas is positively correlated[3]. We would like to choose priors or regularizers that lead to posterior densities or point estimates that take into account this information. Lasso (Tibshirani, 1996) is known to perform poorly in models where there are strong correlations between the regression coefficients, that is, from a group of highly correlated coefficients it chooses a few and suppresses the rest. A way to remedy this drawback is to use the elastic net (Zou and Hastie, 2005). One can go even further and use the group lasso (Meier et al., 2008) which exhibits similar properties as lasso, but the sparsity is represented at a (predefined) group level. However, in many cases it is too restrictive to pre-define the groups. In this chapter, we will define a novel sparsity inducing prior density that allows us to encode prior correlations between the parameters' magnitudes and yields posterior densities that allow us to assess the relevance of the regression coefficients. We apply it in the linear and logistic regression setting.

The chapter is structured in the following way. In Section 4.2 we introduce the notation for the linear regression and logistic regression models followed by a brief discussion in Section 4.3 on the sparsity inducing univariate priors in the MAP and Bayesian framework. In Section 4.3.2, we introduce a sparsity inducing multivariate prior that is defined as a hierarchical scale mixture distribution. In Section 4.4, we present the details of expectation propagation (EP) for approximating the posterior density in the linear and logistic regression models. In Section 4.5, we demonstrate the usefulness of this prior in MEG source localization (Section 4.5.1) and multivariate fMRI analysis (Section 4.5.2) problems.

[3] In the following, we associate the level of activation to the magnitude of the regression coefficients.

4.2 Probabilistic regression and classification with the latent linear model

We consider the linear regression and logistic regression models. The former is used to model continuous observations $\boldsymbol{y} \in \mathbb{R}^m$ while the latter is suited to model binary observations typically encoded by $\boldsymbol{y} \in \{-1, 1\}^m$. The common modeling aspect in these models is that the latent variables $\boldsymbol{f} = (f_1, \ldots, f_m)^T$ defining the parameters of the observation model depend linearly on the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ and that the observations y_i are identically and independently distributed given the latent variables $\boldsymbol{f}$. The dependence is defined through the *design* matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$. The linear regression model is defined formally as

$$y_i \sim \mathcal{N}(f_i, v), \text{ with } \quad \boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta},$$

where v is the variance of the noise. The logistic regression model is defined as

$$y_i \sim \text{Binomial}\left(\frac{e^{f_i}}{1+e^{f_i}}\right), \text{ with } \quad \boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}.$$

The corresponding likelihoods can be written as

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, v) = N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, v\boldsymbol{I})$$

and

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) &= \prod_i \sigma(\boldsymbol{x_i}\boldsymbol{\beta})^{(1+y_i)/2}[1-\sigma(\boldsymbol{x_i}\boldsymbol{\beta})]^{(1-y_i)/2} \\ &= \prod_i \sigma(y_i \boldsymbol{x_i}\boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{x}_i$ is the i^{th} row of $\boldsymbol{X}$ and $\sigma(z) = e^z(1+e^z)^{-1}$. Both likelihoods are log-concave. When $m < n$ the models are called under-determined, because the likelihoods possess a convex manifold of global maxima. In the following, we define a sparsity inducing multivariate prior, which we will use together with these likelihoods to define the corresponding Bayesian models. We start out with a brief overview of the univariate sparsity inducing priors and then proceed to define a multivariate sparsity inducing prior distribution.

4.3 Sparsity inducing priors

If one would want to put in words the properties of the desired values for the regression coefficients, it would be something like: zero with probability p_0 or anything else with some given large average magnitude. In the language of probability, this would correspond to $p(\beta) = p_0\delta_0(\beta) + (1-p_0)N(\beta|0, v_p)$, where v_p is a large positive real. This prior is called the spike and slab prior (e.g. George and McCullogh, 1993). It can be relaxed to the Gaussian mixture prior $p(\beta) = p_0 N(\beta|0, v_0) + (1-p_0)N(\beta|0, v_p)$, where

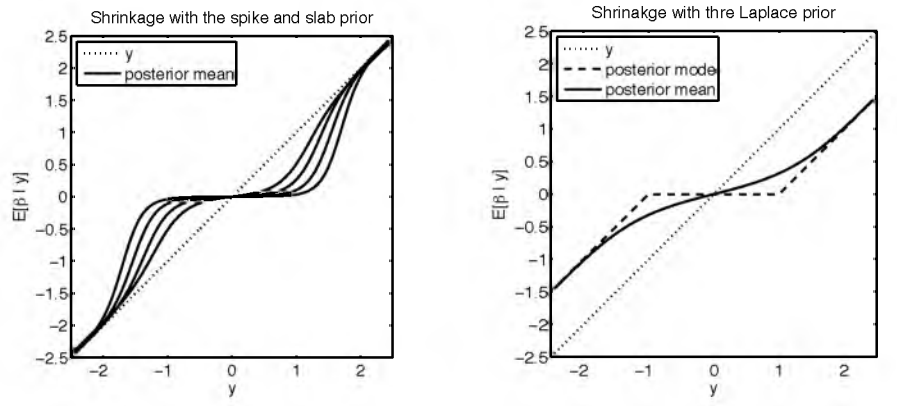

Figure 4.1: The posterior mean values when using a double exponential $e^{-|\beta|/\sqrt{\theta}}/2\sqrt{\theta}$ (right) and a spike and slab $p_0\delta_0(\beta) + (1 - p_0)N(\beta|0, v_p)$ (left) prior with Gaussian likelihood $N(y|, \beta, v)$. The parameters have been set to $v = 0.25$, $\sqrt{\theta} = 0.25$ and $v_p \in \{10, 10^2, 10^3, 10^4\}$ with $p_0 = 0.5$.

v_0 is a small positive real.

It common to characterize sparsity inducing priors by their shrinkage profiles. Let y be an observation of β with Gaussian noise having variance v. Plotting the posterior mean value $\mathrm{E}[\beta|y]$ against y can give us an insight into the shrinkage properties of the prior $p(\beta)$. In the case of the Gaussian mixture prior, computing the posterior mean value and taking the limit $v_0 \to 0$, we get

$$\mathrm{E}[\beta|y] = y\frac{v_p}{v + v_p}\left(1 + \frac{p_0}{1 - p_0}\frac{N(y|0, v)}{N(y|0, v + v_p)}\right)^{-1}.$$

The left panel of Figure 4.1 shows the posterior mean value against the observed value y and it shows that the effect is indeed what we expect: small values of y are shrunk towards zero, while larger values are kept almost unchanged.

Although the spike and slab prior has all the desired properties, it can make approximate inference hard. The main reason is that the log-posterior is multimodal and the current approximate inference methods often fail to converge for such models. A class of prior densities for which approximate inference methods are relatively easy to operate both in the linear regression and logistic regression setting are the log-concave densities (Seeger, 2008). A commonly used log-concave sparsity inducing prior is the double exponential or Laplace prior $L(\beta|\theta) = e^{-|\beta|/\sqrt{\theta}}/(2\sqrt{\theta})$. This prior originates from the regularization framework (Tibshirani, 1996) and thus, it typically works well in the MAP setting. Although due to the bias it introduces, it has a less desirable shrinkage profile than the spike and slap prior, it also has the advantage that approximate inference methods like EP are easier to implement than for the spike and slab prior. In the regularization setting, it is relatively easy to motivate its purpose as a regularizer. It can be viewed as the relaxation of the L_0 norm, that is, the norm counting the non-zero coefficients of a vec-

tor, or it can be used to formulate an upper bound constraint on the coefficient's average absolute value $(\sum_j |\beta_j|)/n$. The right panel of Figure 4.1 shows the double exponential prior's effect on a univariate Gaussian observation y. In the MAP setting, it basically thresholds the posterior mean values at $v/\sqrt{\theta}$ and it creates a constant bias of $v/\sqrt{\theta}$ for all $|y| > v/\sqrt{\theta}$. When applied in Bayesian inference, the thresholding effect is smoothed.

It is common to consider factorizing priors $p(\boldsymbol{\beta}) = \prod_j p(\beta_j)$. However, as mentioned above, in many practical applications there is prior knowledge about the dependence between the variables β_j that should be expressed by the prior. In our case the dependence is correlation in magnitudes.

In the following sections we define a multivariate scale mixture distribution that is based on the univariate double exponential distribution and it correlates the magnitudes of the regression parameters. The distribution is similar in spirit to the multivariate scale mixture distribution defined in Lyu and Simoncelli (2006).

4.3.1 Scale mixture distributions

It has been shown (e.g. Andrews and Mallows, 1974) that when the random variable β has a symmetric probability density such that $(-1)^k \frac{d^k}{d\beta^k} p(\sqrt{\beta}) > 0$ for $\beta \geq 0$ then it can be written as the ratio z/s of a standard normal random variable z and a non-zero random variable s with a density $\phi(s)$ depending on $p(\beta)$. In other words

$$p(\beta) = \int_0^\infty ds \phi(s) s N(s\beta|0,1). \tag{4.1}$$

The relation between $p(\beta)$ and $\phi(s)$ can be found by using the Laplace transform w.r.t. $p(\sqrt{\beta})$ and $\phi(\sqrt{s})$. This representation is called the scale mixture representation and is a useful form to represent symmetric distributions obeying the above mentioned mild constraints, or to define new distributions. When using it for defining distributions, its usefulness is in its hierarchical representation which allows a good control over the placement of the distribution's mass.

4.3.2 A multivariate sparsity inducing scale mixture prior

The double exponential distribution $L(\beta|\theta)$ can be written in the scale mixture form with the help of the exponential distribution $\mathcal{E}(\gamma|\theta) = e^{-\gamma/\sqrt{\theta}}/\sqrt{\theta}$. By a change of variables, the scale mixture form in (4.1) for the double exponential distribution is given by

$$L(\beta|\theta) = \int_0^\infty d\gamma \, \mathcal{E}(\gamma|2\theta) N(\beta|0,\gamma). \tag{4.2}$$

The distribution $L(\beta|\theta)$ is zero centered and the variance of β is 2θ. Other choices of $\phi(s)$ lead to sparsity inducing univariate priors like the horseshoe prior (Carvalho et al., 2009), the Student-t prior (Tipping, 2001), the Gaussian mixture or spike and slab prior (e.g. George and McCullogh, 1993) and many more. Carvalho et al. (2008) give a detailed comparison of the shrinkage properties of these priors.

The multivariate double exponential distribution $L_{\text{mv}}(\boldsymbol{\beta})$ can be defined with the help of its characteristic function $\phi(\boldsymbol{t}) = (1 + \boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t})^{-1}$ (Kotz et al., 2001). However, when using it in a Bayesian setting, the scale mixture formulation

$$L_{\text{mv}}(\boldsymbol{\beta}|\boldsymbol{\Sigma}) = \int_0^\infty d\gamma \, \mathcal{E}(\gamma|2) N(\boldsymbol{\beta}|\mathbf{0}, \gamma\boldsymbol{\Sigma}),$$

is more suitable, because of its hierarchical representation. The random vector $\boldsymbol{\beta} \sim L_{\text{mv}}$ has zero mean and covariance $\text{V}[\boldsymbol{\beta}] = 2\boldsymbol{\Sigma}$. Eltoft et al. (2006) showed that the analytical form of $L_{\text{mv}}(\boldsymbol{\beta}|\boldsymbol{\Sigma})$ is

$$L_{\text{mv}}(\boldsymbol{\beta}|\boldsymbol{\Sigma}) = (2\pi)^{-n/2} \left(\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}}\right)^{1-n/2} \mathcal{K}_{n/2-1}\left(\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}}\right),$$

where $\mathcal{K}_{n/2-1}$ is the modified Bessel function of the second kind and order $n/2 - 1$. Eltoft et al. (2006) also showed that $L_{\text{mv}}(\boldsymbol{\beta}|\boldsymbol{\Sigma})$ has heavier tails than a Gaussian. The above form shows that the multivariate Laplace distributions penalizes high values of $\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}$, thus the shrinking effect acts on the whole Mahalanobis norm. Unfortunately, this is not the property we desire. Our goal is to define a prior where the variables β_j interact, but the shrinkage does not act "globally" on $\boldsymbol{\beta}$. In order to deal with this problem, Lyu and Simoncelli (2006) introduce a scaling parameter for every variable β_j and define a Gaussian multivariate prior on the log of the scaling parameters. The model was successfully applied to image de-noising by using the MAP approach.

Following the idea of using a different scaling parameter for each β_j, we define a multivariate generalization of the double exponential distribution in a similar spirit as Lyu and Simoncelli (2006). We use a multivariate exponential prior on the scaling parameters. The multivariate exponential prior in Longford (1990) is based on the idea that an univariate exponential random variable can be represented as the sum of squares of two independent univariate normally distributed random variables. These normally distributed random variables can then be coupled in two identical multivariate normal distributions.

Using the arguments mentioned above, the scale mixture form of the double exponential distribution in (4.2) can be written as

$$L(\beta|\theta) = \int du dv \, N(u|0, \theta) N(v|0, \theta) N(\beta|0, u^2 + v^2)$$

and thus, the factorizing double exponential prior $p(\boldsymbol{\beta}|\theta)$ has the form

$$p(\boldsymbol{\beta}|\theta) = \int d\boldsymbol{u} d\boldsymbol{v} \, N(\boldsymbol{u}|\mathbf{0}, \theta \boldsymbol{I}) N(\boldsymbol{v}|\mathbf{0}, \theta \boldsymbol{I}) \prod_j N(\beta_j|0, u_j^2 + v_j^2).$$

By using a multivariate exponential prior, we generalize this formulation and define the hierarchical distribution

$$L(\boldsymbol{\beta}|\boldsymbol{\Sigma}) \equiv \int d\boldsymbol{u} d\boldsymbol{v} \, N(\boldsymbol{u}|\mathbf{0}, \boldsymbol{\Sigma}) N(\boldsymbol{v}|\mathbf{0}, \boldsymbol{\Sigma}) \prod_j N(\beta_j|0, u_j^2 + v_j^2).$$

The characteristic function of this distribution is $\phi(\boldsymbol{t}) = |\boldsymbol{I} + \boldsymbol{\Sigma}\mathrm{diag}(\boldsymbol{t}^2)|^{-1}$, showing that the two distributions $L_{\mathrm{mv}}(\boldsymbol{\beta}|\boldsymbol{\Sigma})$ and $L(\boldsymbol{\beta}|\boldsymbol{\Sigma})$ have significantly different properties. Intuitively, the distribution $L(\boldsymbol{\beta}|\boldsymbol{\Sigma})$ correlates the magnitudes of the variables β_j. It can indeed be verified that the variables β_j are uncorrelated, as opposed to $L_{\mathrm{mv}}(\boldsymbol{\beta}|\boldsymbol{\Sigma})$ where they are correlated according to $2\boldsymbol{\Sigma}$. The β_js are zero-mean and their marginal variance is $\mathrm{E}\left[\beta_j^2\right] = 2\Sigma_{jj}$. The covariance between β_i^2 and β_j^2 is $\mathrm{Cov}\left[\beta_i^2, \beta_j^2\right] = 2\Sigma_{ij}^2$. Thus by using a different scaling parameter for each β_j and imposing a multivariate exponential prior on these scaling parameters, we have defined a multivariate density which leaves the variables β_j (marginally) uncorrelated, but correlates their magnitudes. When no correlation is present, in other words when $\Sigma_{ij} = 0$, it boils down to the product of independent double exponential priors.

In the following, we show how EP can be applied in an efficient way to perform approximate inference in the linear and the logistic regression model using the scale mixture prior $L(\boldsymbol{\beta}|\boldsymbol{\Sigma})$.

4.4 Approximate Inference

In the linear regression case, the joint posterior density of the variables $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \propto N(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{\Sigma}) N(\boldsymbol{v}|\boldsymbol{0}, \boldsymbol{\Sigma}) \prod_j N(\beta_j|0, u_j^2 + v_j^2) N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, v\boldsymbol{I}) \quad (4.3)$$

while in the logistic regression case it is

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \propto N(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{\Sigma}) N(\boldsymbol{v}|\boldsymbol{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{n} N(\beta_j|0, u_j^2 + v_j^2) \prod_{i=1}^{m} \sigma(y_i \boldsymbol{x}_i^T \boldsymbol{\beta}). \quad (4.4)$$

In order to simplify notation, we omit the posterior densities' dependence on $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$. We will approximate both distributions with Gaussian distributions $q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v})$ using EP, but before starting any approximation we take a close look at (4.3) and (4.4) to see if there are some properties of the posterior densities that we can make use of. Both (4.3) and (4.4) are invariant w.r.t. sign changes of $\boldsymbol{u}$ and $\boldsymbol{v}$. Assuming that we could integrate out $\boldsymbol{\beta}$, the remaining $p(\boldsymbol{u}, \boldsymbol{v})$s would also possess this symmetry property, thus $\mathrm{E}_{p(\boldsymbol{u},\boldsymbol{v})}\left[(\boldsymbol{u}^T, \boldsymbol{v}^T)\right] = \boldsymbol{0}$ and $\mathrm{E}_{p(\boldsymbol{u},\boldsymbol{v})}\left[\boldsymbol{u}\boldsymbol{v}^T\right] = \boldsymbol{0}$. One can also check that $p(\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{v}) = p(\boldsymbol{\beta}|\boldsymbol{u}^2 + \boldsymbol{v}^2)$ both for (4.3) and (4.4). This, together with the symmetry of the $p(\boldsymbol{u}, \boldsymbol{v})$s, implies that $\mathrm{E}_{p(\boldsymbol{u},\boldsymbol{v})}\left[\mathrm{E}_{p(\boldsymbol{\beta}|\boldsymbol{u},\boldsymbol{v})}[\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{v}]\, \boldsymbol{u}^T\right] = \boldsymbol{0}$. These properties simplify significantly the application of EP, because we are approximating the first two moments of random vectors that have block-diagonal covariance structures.

Let $\tilde{t}_{\mathrm{N},j}(\beta_j, u_j, v_j)$ and $\tilde{t}_{\sigma,i}(\boldsymbol{\beta})$ be the Gaussian term approximations corresponding to the non-Gaussian terms $t_{\mathrm{N},j}(\beta_j, u_j, v_j) = N(\beta_j, 0, u_j^2 + v_j^2)$ and $t_{\sigma,i}(\boldsymbol{\beta}) = \sigma(y_i \boldsymbol{x}_i \boldsymbol{\beta})$ respectively. Since the logistic regression case (4.4) has both $t_{\mathrm{N},j}$ and $t_{\sigma,i}$ terms, we start with presenting EP for approximating (4.3) and then proceed to approximate (4.4), where we only have to sort out the approximations for the terms $t_{\sigma,i}$ and omit the Gaussian term $N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, v^{-1}\boldsymbol{I})$. In order to unify the notation, we introduce the variable $\boldsymbol{z}^T = (\boldsymbol{\beta}^T, \boldsymbol{u}^T, \boldsymbol{v}^T)$. Both types of terms $t_{\mathrm{N},j}$ and $t_{\sigma,i}$ can be viewed as depending on a linear transformation of the variables $\boldsymbol{U}_{\mathrm{N},j}\boldsymbol{z}$ and $\boldsymbol{U}_{\sigma,i}\boldsymbol{z}$. In case of the terms $t_{\mathrm{N},j}$,

$\boldsymbol{U}_{\mathrm{N},j}\boldsymbol{z} \in \mathbb{R}^{3\times 3n}$ can be written as $\boldsymbol{U}_{\mathrm{N},j} = (\boldsymbol{e}_j, \boldsymbol{e}_{n+j}, \boldsymbol{e}_{2n+j})^T$, where $\boldsymbol{e}_k$ is the k^{th} unit vector in $\mathbb{R}^{3n}$, while in case of the terms $t_{\sigma,i}$, $\boldsymbol{U}_{\sigma,j}\boldsymbol{z} \in \mathbb{R}^{1\times 3n}$, can be written as $\boldsymbol{U}_{\sigma,i} = (y_i\boldsymbol{x}_i^T, \mathbf{0}^T, \mathbf{0}^T)$, where $\mathbf{0} \in \mathbb{R}^n$. A generic presentation of EP for the case when the terms depend on linear transformations of the variables is given in Section A.4 of the Appendix, thus we are left with identifying the transformations $\boldsymbol{U}_{\mathrm{N},j}$ and $\boldsymbol{U}_{\sigma,i}$, and performing the corresponding steps. In the following, we use the notation $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ and detail these steps both in the linear and the logistic regression case.

4.4.1 The linear regression model

In case of the linear regression model we only have to perform the EP updates for the $\tilde{t}_{\mathrm{N},j}$ terms. We choose p_0 as $p_0(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}) \propto N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, v\boldsymbol{I})N(\boldsymbol{u}|\mathbf{0}, \boldsymbol{Q}^{-1})N(\boldsymbol{v}|\mathbf{0}, \boldsymbol{Q}^{-1})$. Since the terms $t_{\mathrm{N},j}$ depend on the variables β_j, u_j and v_j, their term approximations $\tilde{t}_{\mathrm{N},j}$ have the form

$$\log \tilde{t}_{\mathrm{N},j}(\beta_j, u_j, v_j) = (\beta_j, u_j, v_j)\tilde{\boldsymbol{h}}^j - \frac{1}{2}(\beta_j, u_j, v_j)\tilde{\boldsymbol{Q}}^j(\beta_j, u_j, v_j)^T, \tag{4.5}$$

and thus the approximating distribution has the form

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}) \propto p_0(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}) \prod_j \tilde{t}_{\mathrm{N},j}(\beta_j, u_j, v_j).$$

The canonical parameters $\tilde{\boldsymbol{h}}$ and $\tilde{\boldsymbol{Q}}$ of q are

$$\tilde{\boldsymbol{h}} = \begin{bmatrix} (\tilde{h}_1^j)_j \\ (\tilde{h}_2^j)_j \\ (\tilde{h}_3^j)_j \end{bmatrix}, \tilde{\boldsymbol{Q}} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X}/v + \mathrm{diag}(\tilde{Q}_{11}^j) & \mathrm{diag}(\tilde{Q}_{12}^j) & \mathrm{diag}(\tilde{Q}_{13}^j) \\ \mathrm{diag}(\tilde{Q}_{21}^j) & \boldsymbol{Q} + \mathrm{diag}(\tilde{Q}_{22}^j) & \mathrm{diag}(\tilde{Q}_{23}^j) \\ \mathrm{diag}\tilde{Q}_{31}^j) & \mathrm{diag}(\tilde{Q}_{32}^j) & \boldsymbol{Q} + \mathrm{diag}(\tilde{Q}_{33}^j) \end{bmatrix}$$

As we have seen earlier, the covariance structure corresponding to (4.3) is block diagonal. It therefore makes sense to initialize the term approximations $\tilde{t}_{\mathrm{N},j}$ to a factorizing form $\tilde{t}_{\mathrm{N},j}(\beta_j, u_j, v_j) = \tilde{t}_{\mathrm{N},j}(\beta_j)\tilde{t}_{\mathrm{N},j}(u_j)\tilde{t}_{\mathrm{N},j}(v_j)$ implying that $\tilde{\boldsymbol{Q}}$ also has a block diagonal structure, that is $\tilde{Q}_{lk}^j = 0$ for all $l \neq k$. In the following, we work out the EP updates for the term approximations $\tilde{t}_{\mathrm{N},j}$ in more detail and show by induction that starting from factorized forms, the $\tilde{t}_{\mathrm{N},j}$s will keep their factorized form throughout the algorithm. Since u_i and v_i play identical roles, both q and the term approximations are symmetric in u_i and v_i, we have $\tilde{h}_2^j = \tilde{h}_3^j$ and $\tilde{Q}_{22}^j = \tilde{Q}_{33}^j$ for all $j = 1, \ldots, n$.

Term updates

Let us assume that the $\tilde{t}_{\mathrm{N},j}$s factorize. Then, due to the block diagonal structure, the densities $q(\beta_j, u_j, v_j)$ and consequently $q^{\setminus j}(\beta_j, u_j, v_j)$ factorize. Now, let the cavity distribution $q^{\setminus j}(\beta_j, u_j, v_j)$ have the parametric form

$$q^{\setminus j}(\beta_j, u_j, v_j) = N(\beta_j|m_j, \tau_j^2)N(u_j|0, \gamma_j^2)N(v_j|0, \gamma_j^2). \tag{4.6}$$

To obtain a new term approximation we need to compute the moments of the *tilted* marginal $q_j(\beta_j, u_j, v_j) = N(\beta_j|0, u_j^2 + v_j^2)^\alpha q^{\setminus j}(\beta_j, u_j, v_j)$, which, by a regrouping of

terms, can be written in the conditional form $q_j(\beta_j, u_j, v_j) = q_j(\beta_j|u_j, v_j)q_j(u_j, v_j)$ with

$$q_j(\beta_j|u_j, v_j) = N\left(\beta_j|\frac{m_j(u_j^2+v_j^2)}{\alpha\tau_j^2+u_j^2+v_j^2}, \frac{\tau_j^2(u_j^2+v_j^2)}{\alpha\tau_j^2+u_j^2+v_j^2}\right) \tag{4.7}$$

$$\begin{aligned} q_j(u_j, v_j) &= (u_j^2+v_j^2)^{(1-\alpha)/2}N(\sqrt{\alpha}m_j|0, \alpha, \alpha\tau_j^2+u_j^2+v_j^2) \\ &\quad \times N(u_j|0, \gamma_j^2)N(v_j|0, \gamma_j^2). \end{aligned} \tag{4.8}$$

An inspection of (4.7) and (4.8) shows that both $q_j(\beta_j|u_j, v_j)$ and $q_j(u_j, v_j)$ depend on u_j and v_j only through $u_j^2+v_j^2$. Therefore, both distributions are invariant under sign changes of u_j and v_j. These observations imply that $\mathrm{E}_{q_j}[u_j] = \mathrm{E}_{q_j}[v_j] = 0$, $\mathrm{E}_{q_j}[u_j v_j] = 0$ and $\mathrm{E}_{q_j}\left[u_j^2\right] = \mathrm{E}_{q_j}\left[v_j^2\right]$. Since $q_j(u_j, v_j)$ can be expressed as a function of $u_j^2+v_j^2$ and the sum of squares of two identically distributed normal variables is exponentially distributed, we can use univariate Gauss-Laguerre numerical quadrature to compute $\mathrm{E}_{q_j}\left[u_j^2+v_j^2\right]$. The symmetry arguments imply that $\mathrm{E}_{q_j}\left[u_j^2\right] = \mathrm{E}_{q_j}\left[v_j^2\right] = \mathrm{E}_{q_j}\left[u_j^2+v_j^2\right]/2$. Now we can proceed to compute the marginal moments of β_j w.r.t. q_j. The marginal mean and variance of β_j can be computed from the averages the mean and variance parameters in (4.7) w.r.t. $q_j(u_j, v_j)$. These averages can be computed again by univariate Gauss-Laguerre numerical quadratures w.r.t. $u_j^2+v_j^2$. Computing the covariances boils down to computing $\mathrm{E}_{q_j}[\beta_j u_j]$. We can write $\mathrm{E}_{q_j}[\beta_j u_j] = \mathrm{E}_{q_j(u_j,v_j)}\left[u_j \mathrm{E}_{q_j(\beta_j|u_j,v_j)}[\beta_j|u_j, v_j]\right]$, which, due to symmetry, is again $\mathrm{E}_{q_j}[\beta_j u_j] = \mathrm{E}_{q_j}[\beta_j v_j] = 0$.

Computing the cavity distribution

In order to compute $q^{\setminus j}(\beta_j, u_j, v_j) = q^{\setminus j}(\beta_j)q^{\setminus j}(u_j)q^{\setminus j}(v_j)$, we need to compute the marginals $q(\beta_j)$, $q(u_j)$ and $q(v_j)$ to form $q^{\setminus j}(\beta_j)$, $q^{\setminus j}(u_j)$ and $q^{\setminus j}(v_j)$. The computation of the marginals boils down to computing the diagonal elements of $(\boldsymbol{Q} + \mathrm{diag}(\tilde{Q}_{11}^j))^{-1}$ (remember that $\tilde{Q}_{11}^j = \tilde{Q}_{22}^j$) and $(\boldsymbol{X}^T\boldsymbol{X}/v + \mathrm{diag}(\tilde{Q}_{11}^j))^{-1}$ and solving the linear system $(\boldsymbol{X}^T\boldsymbol{X}/v + \mathrm{diag}(\tilde{Q}_{11}^j))\boldsymbol{m} = (\tilde{h}_1^j)_j$ (remember that due to symmetry $\tilde{h}_2^j = \tilde{h}_3^j = 0$). To compute the diagonal elements of $(\boldsymbol{Q} + \mathrm{diag}(\tilde{Q}_{22}^j))^{-1}$, we use the Takahashi equations, as in Section 3.4.6, and make use of the sparsity of $\boldsymbol{Q}$ (when it is the case). When computing the diagonal elements of $(\boldsymbol{X}^T\boldsymbol{X}/v + \mathrm{diag}(\tilde{Q}_{11}^j))^{-1}$ and solving the system $(\boldsymbol{X}^T\boldsymbol{X}/v + \mathrm{diag}(\tilde{Q}_{11}^j))\boldsymbol{m} = (\tilde{h}_1^j)_j$, we either perform the inversion directly (by using Cholesky factorization), when $n \leq m$, or use the Woodbury formula (e.g. Golub and van Loan, 1996) followed by a Cholesky factorization of $\boldsymbol{X}\mathrm{diag}(\tilde{Q}_{11}^j)^{-1}\boldsymbol{X}^T + v\boldsymbol{I}$, when $n > m$.

4.4.2 The logistic regression model

In case of the logistic regression model (4.4) the prior p_0 is chosen as $p_0(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}) = N(\boldsymbol{u}|0, \boldsymbol{Q}^{-1})N(\boldsymbol{v}|0, \boldsymbol{Q}^{-1})$ and the terms $t_{\sigma,i}(\boldsymbol{\beta}) = \sigma(y_i\boldsymbol{x}_i\boldsymbol{\beta})$ depend only on $\boldsymbol{\beta}$. The term approximations $\tilde{t}_{\mathrm{N},j}(\beta_j, u_j, v_j)$ are the same as in (4.5) while the term approximations $\tilde{t}_{\sigma,i}(\boldsymbol{\beta})$ have the form

$$\log \tilde{t}_{\sigma,i}(\boldsymbol{\beta}) = \boldsymbol{\beta}^T[y_i\boldsymbol{x}_i]^T\tilde{h}^i - \frac{1}{2}\boldsymbol{\beta}^T[\boldsymbol{x}_i]^T\tilde{Q}^j[\boldsymbol{x}_i]\boldsymbol{\beta}$$

steps \ variables	$\boldsymbol{\beta}$		$\boldsymbol{u}$ and $\boldsymbol{v}$	
model	linear	logistic	linear	logistic
term updates	$n \times n_{\text{grid}}$	$(n+m) \times n_{\text{grid}}$	$n \times n_{\text{grid}}$	$n \times n_{\text{grid}}$
cavities	$n \times \min(m,n)^2$	$(n+m) \times \min(m,n)^2$	$n_C(\boldsymbol{Q}) + n_T(\boldsymbol{L})$	$n_C(\boldsymbol{Q}) + n_T(\boldsymbol{L})$

Table 4.1: Computational complexities of the EP steps for the linear and logistic regression models with the sparsity inducing prior $L(\boldsymbol{\beta}|\boldsymbol{Q}^{-1})$. $\boldsymbol{L}$ is the Cholesky factor $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$, $n_C(\boldsymbol{Q})$ is the complexity of the sparse Cholesky factorization and $n_T(\boldsymbol{L})$ is the complexity of solving the Takahashi equations given $\boldsymbol{L}$. Both scale roughly with $nnzeros(\boldsymbol{Q})^2/n$.

with $y_i^2 = 1$ and thus, the canonical parameters $\tilde{\boldsymbol{h}}$ and $\tilde{\boldsymbol{Q}}$ of approximation q are

$$\tilde{\boldsymbol{h}} = \begin{bmatrix} \boldsymbol{X}^T[\boldsymbol{y}(\tilde{h}^i)_i] + (\tilde{h}_1^j)_j \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

$$\tilde{\boldsymbol{Q}} = \begin{bmatrix} \boldsymbol{X}^T\text{diag}((\tilde{Q}^i)_i)\boldsymbol{X} + \text{diag}(\tilde{Q}_{11}^j) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q} + \text{diag}(\tilde{Q}_{22}^j) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q} + \text{diag}(\tilde{Q}_{33}^j) \end{bmatrix}.$$

The details of the approximate term updates $\tilde{t}_{\text{N},j}(\beta_j, u_i, v_j)$ are exactly the same as in case of the linear regression model, therefore, we only have to work out the details for the $\tilde{t}_{\sigma,i}(\boldsymbol{\beta})$ terms.

Term updates

Let $q^{\backslash i}(\boldsymbol{\beta}) \propto q(\boldsymbol{\beta})/\tilde{t}_{\sigma,i}(\boldsymbol{\beta})^\alpha$ have the parametric form $q^{\backslash i}(\boldsymbol{\beta}) = N(\boldsymbol{\beta}|\boldsymbol{m}^{\backslash i}, \boldsymbol{V}^{\backslash i})$, then according to Section A.4 of the Appendix, in order to update the term $\tilde{t}_{\sigma,i}(\boldsymbol{\beta})$, we have to compute the moments of $q_i(s) \propto \sigma(s)^\alpha N(s|y_i\boldsymbol{x}_i\boldsymbol{m}^{\backslash i}, \boldsymbol{x}_i\boldsymbol{V}^{\backslash i}\boldsymbol{x}_i^T)$. We can compute the moments by univariate numerical quadrature and update the terms accordingly.

Computing the cavity distribution

The cavity distributions to be computed are the following: (1) $q^{\backslash i}(\boldsymbol{\beta}) \propto q(\boldsymbol{\beta})/\tilde{t}_{\sigma,i}(\boldsymbol{\beta})^\alpha$ for the terms $\tilde{t}_{\sigma,i}$ and (2) $q^{\backslash j}(\beta_j), q^{\backslash j}(u_j)$ and $q^{\backslash j}(v_j)$ for the terms $\tilde{t}_{\text{N},j}$. The marginals $q_j(u_j)$, $q_j(v_j)$ and $q_j(\beta_j)$ can be computed in the same way as in case of the linear regression model. As shown in Section A.4 of the Appendix, computing the cavity $q^{\backslash i}(\boldsymbol{\beta})$ boils down to computing the projection of $q(\boldsymbol{\beta})$'s moment parameters into $y_i\boldsymbol{x}_i$ for all $i = 1, \ldots, m$, that is, to the computation of the quadratic forms $\boldsymbol{x}_i[\boldsymbol{X}^T\text{diag}((\tilde{Q}^i)_i)\boldsymbol{X} + \text{diag}(\tilde{Q}_{11}^j)]^{-1}\boldsymbol{x}_i^T$ and $y_i\boldsymbol{x}_i[\boldsymbol{X}^T\text{diag}((\tilde{Q}^i)_i)\boldsymbol{X} + \text{diag}(\tilde{Q}_{11}^j)]^{-1}[\boldsymbol{X}^T[\boldsymbol{y}(\tilde{h}^i)_i] + (\tilde{h}_1^j)_j]$. Similarly to the linear regression case, we can perform the computations by using a Cholesky factorization, when $n \leq m$, or using the Woodbury formula and thereafter performing a Cholesky factorization of $\boldsymbol{X}\text{diag}(\tilde{Q}_{11}^j)^{-1}\boldsymbol{X}^T + \text{diag}((\tilde{Q}^i)_i)^{-1}$, when $n > m$.

4.4.3 Computational complexities

The computational complexities are summarized in Table 4.1. The computational bottleneck of the inference algorithm is computing the parameters of the cavity distribution. When $n < m$, the computational complexity of computing the marginals moments in $q(\boldsymbol{\beta})$ or their projections is dominated by the (full) Cholesky factorization for both models (see Sections 4.4.1 and 4.4.2). When $m < n$, we use the Woodbury formula and the computational complexity is dominated by the computation of the diagonal of $\boldsymbol{X}^T(\boldsymbol{X}\mathrm{diag}(\tilde{Q}^j)^{-1}\boldsymbol{X} + v\boldsymbol{I})^{-1}\boldsymbol{X}$ or $\boldsymbol{X}^T(\boldsymbol{X}\mathrm{diag}((\tilde{Q}^j)_j)^{-1}\boldsymbol{X} + \mathrm{diag}((\tilde{Q}^i)_i)^{-1})^{-1}\boldsymbol{X}$. Both scale with nm^2. The computational complexity of computing the marginal moments in $q(\boldsymbol{u})$ scales roughly with $nnzeros(\boldsymbol{Q})^2/n$ (see Section A.2 of the Appendix). We used the AMD re-ordering algorithm (Amestoy et al., 1996), the MATLAB implementation of the sparse Cholesky factorization and implemented the algorithm for solving the Takahashi equations in C. In the models studied in Sections 4.5.1 and 4.5.2, the computational time was dominated by the latter.

4.5 Examples

The approximate inference methods presented above were be successfully applied for identifying the activation of brain areas in task-related MEG and fMRI experiments.

4.5.1 Source localization

We apply the linear regression model to a Bayesian source localization problem. Our approach of using the multivariate scale mixture prior is illustrated by an experiment based on the mismatch negativity paradigm for which MEG data and a structural MRI have been acquired.

Problem description

We consider the problem of identifying regions of brain activity when using MEG measurements in a certain experimental setting presented in the next section. The sensor readings $\boldsymbol{y} \in \mathbb{R}^m$ and the source currents $\boldsymbol{\beta} \in \mathbb{R}^n$ are assumed to be related as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is called the *lead field matrix* and $\boldsymbol{\epsilon}$ is a vector of normally and independently distributed sensor noises with variance v. Localizing distributed sources is an ill-posed inverse problem that only admits a unique solution when additional constrains are defined on the source currents which in the Bayesian setting take the form of priors (Wipf and Nagarajan, 2009). Typical choices are the Gaussian and the double exponential prior. The corresponding MAP solutions are the minimum norm and the minimum current estimates (Haufe et al., 2008). The minimum norm estimate produces spatially smooth solutions, but is known to suffer from depth bias and smearing of nearby sources, while the minimum current estimate leads to source estimates that may be scattered too much throughout the brain volume. In order to overcome these problems, we use the sparsifying scale mixture prior introduced in Section 4.3.2 in the context of Bayesian linear regression. Instead of the MAP paradigm, we adopt an approximate Bayesian approach and compute a Gaussian approximation of the posterior source currents $\boldsymbol{\beta}$ and the auxiliary

variables $\boldsymbol{u}$ and $\boldsymbol{v}$. The marginal variance of the auxiliary variables $\boldsymbol{u}$ and $\boldsymbol{v}$ will allow us to define smooth importance maps for identifying regions of brain activity.

Experimental setting

We use a dataset obtained from a mismatch negativity (MMN) experiment (Garrido et al., 2009). The MMN is the negative component of the difference between responses to normal and deviant stimuli within an oddball paradigm that peaks around 150 ms after stimulus onset. In our experiment, the subject had to listen to normal (500 Hz) and deviant (550 Hz) tones, presented for 70 ms. Normal tones occurred 80% of the time, whereas deviants occurred 20% of the time. A total of 600 trials was acquired.

The data was recorded with a CTF MEG System (VSM MedTech Ltd., Coquitlam, British Columbia, Canada), which provides whole-head coverage using 275 DC SQUID axial gradiometers. A realistically shaped volume conduction model was constructed based on the individual's structural MRI. The brain volume was discretized to a grid with a 0.75 cm resolution and the lead field matrix $\boldsymbol{X}$ was calculated for each of the 3863 grid points according to the head position in the system and the forward model. The lead field matrix is defined for the three x, y, and z orientations in each of the source locations and was normalized to correct for depth bias. Consequently, $\boldsymbol{X}$ is of size 275×11589. The 275×1 observation vector $\boldsymbol{y}$ was rescaled to prevent issues with numerical precision.

Specifying the prior

In the context of source identification, we define the prior on the latent variables $\boldsymbol{u}$ and $\boldsymbol{v}$ as follows. For each source current k, there are three corresponding component variables β_{k_x}, β_{k_y} and β_{k_z}. These components form the vector $\boldsymbol{\beta}$ (orientations x, y, and z for all $k = 1, \ldots, n$). To all variables in $\boldsymbol{\beta}$ we can assign a corresponding variable both in $\boldsymbol{u}$ and $\boldsymbol{v}$. Defining $\boldsymbol{Q}$ correponds to specifying the interaction strengths between the variables u_{k_w}, $k = 1, \ldots, n$, $w \in \{x, y, z\}$. As mentioned earlier, our prior assumption is that neighboring sources should have similar magnitudes. For this, we need to couple the triples $(u_{k_x}, u_{k_y}, u_{k_z})$ corresponding to the neighboring sources. We choose to couple the source components individually, that is, we assume that there is no a priori coupling between u_{k_a} and u_{k_b} for $a \neq b$. With these assumptions, we define the prior on $\boldsymbol{u}$ in the spirit of

$$p(\boldsymbol{u}) \propto \exp\left\{ -\frac{1}{2} \sum_{k} \sum_{w \in \{x,y,z\}} u_{k_w}^2 - \frac{1}{2} s \sum_{i \sim j} \sum_{w \in \{x,y,z\}} (u_{i_w} - u_{j_w})^2 \right\}, \tag{4.9}$$

where $i \sim j$ denotes our choice for the neighborhood structure in the brain volume. However, for a well parameterized definition of $\boldsymbol{Q}$, a bit more care is needed. Let $\boldsymbol{W}$ be the precision matrix in (4.9), that is, $W_{i_x,j_x} = W_{i_y,j_y} = W_{i_z,j_z} = -s$ for $i \sim j$, $W_{k_w,k_w} = 1 + s \sum_{j \sim k}$ for all $k = 1, \ldots, n$, $w \in \{x, y, z\}$ and zero for all other entries. Then, we choose the prior precision matrix for $\boldsymbol{u}$ and $\boldsymbol{v}$ as

$$\boldsymbol{Q} = \theta^{-1} \mathrm{diag}(\boldsymbol{W}(s)^{-1})^{1/2} \boldsymbol{W}(s) \mathrm{diag}(\boldsymbol{W}(s)^{-1})^{1/2}.$$

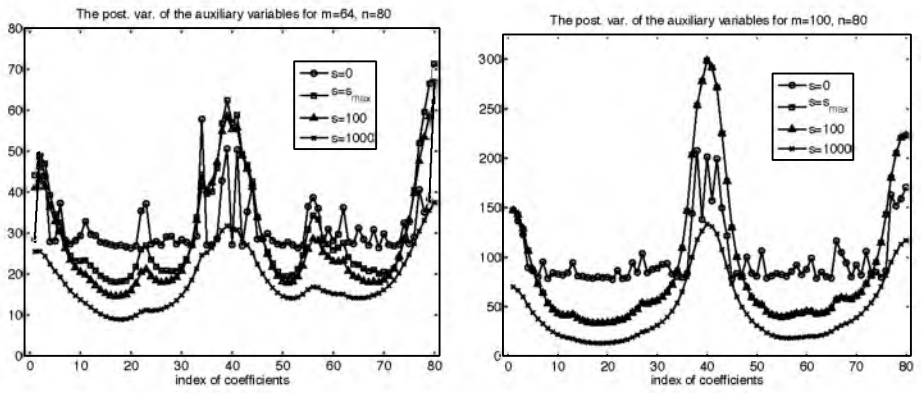

Figure 4.2: The panels of the figure show the approximate posterior variance of the auxiliary variables $u_1, \ldots, u_n$ for different settings of the logistic regression model with the θ parameter set to the value that maximizes EP's approximate evidence. The data was generated using a $\boldsymbol{\beta}$ parameter vector consisting of 4 ones, 32 zeros, 8 minus ones, 32 zeros and 4 ones.

This formulation allows us to define a prior with only two parameters and to control the covariance $\mathrm{diag}(\boldsymbol{W}^{-1}(s))^{-1/2}\boldsymbol{W}(s)^{-1}\mathrm{diag}(\boldsymbol{W}^{-1}(s))^{-1/2}$ and variance θ in $\boldsymbol{Q}^{-1}$ independently. We can use the parameter s to control the prior "smoothness" and θ to control the prior variance. For varying the coupling strength s, we use as a guiding principle the correlation structure it results in. For example, when $s = 10$ the correlations between the x components of two neighboring sources is 0.78 and it decreases with the distance in the grid.

The panels of Figure 4.2 show the approximate posterior variance of the auxiliary variables in $\boldsymbol{u}$ and $\boldsymbol{v}$ for different settings of the logistic regression model for a low dimensional toy model. The prior on $\boldsymbol{u}$ and $\boldsymbol{v}$ was defined to couple the variables u_{j+1} and u_j and v_{j+1} and v_j respectively, that is, the grid was chosen as a one-dimensional. The plots in the figure show that the approximate posterior variances of the auxiliary variables can be used to assess the importance of coefficients and how s controls the smoothness of these values.

Figure 4.3 demonstrates how a chosen coupling leads to a particular structure in $\boldsymbol{Q}$. The irregularities in $\boldsymbol{Q}$ are caused by the structure of the imaged brain volume. The figure also shows the computational bottleneck of our algorithm, that is, the computation of the diagonal elements of $\boldsymbol{Q}^{-1}$ by means of the Takahashi equation. (Remember that $\boldsymbol{Q}$ has the same structure as the block of $\tilde{\boldsymbol{Q}}$ corresponding to $\boldsymbol{u}$.) The block diagonal structure of $\boldsymbol{L}$ arises from a reordering of rows and columns using, for instance, the AMD algorithm (Amestoy et al., 1996). The correlation matrix $\boldsymbol{C}$ shows the correlations between the sources induced by the structure of $\boldsymbol{Q}$. Zeros in the correlation matrix arise from the independence between source orientations x, y, and z.

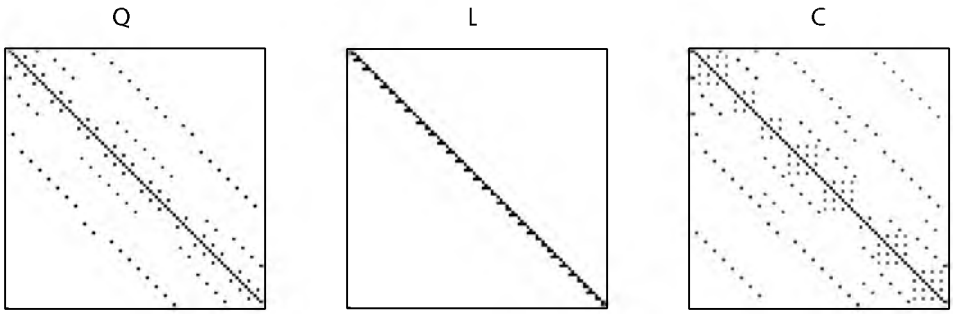

Figure 4.3: Spatial coupling leads to the normalized precision matrix $\boldsymbol{Q}$ with coupling of neighboring source orientations in the x, y, and z directions. The (reordered) matrix $\boldsymbol{L}$ is obtained from the Cholesky decomposition of $\boldsymbol{Q}$. The correlation matrix $\boldsymbol{C}$ shows the correlations between the source orientations. For the purpose of demonstration, we show matrices using a very coarse discretization of the brain volume.

Results

The model we defined has three parameters: (1) the variance v of the noise ϵ_i, (2) the prior variance parameter θ and (3) the coupling strength s. We consider the noise variance and the prior variance fixed to values estimated by the L-curve criterion (Hansen, 1998) and we analyze the model's behavior for varying coupling strength values. When choosing $s = 0$, we arrive at the model with facrotrizing double exponential prior on β_j, whereas for $s \neq 0$ we expect to obtain smooth importance maps. We define the importance of voxel k as the variance of the variables in $\boldsymbol{u}$. Since only sources with non-zero contributions should have high variance, this measure can be used to indicate the relevance of a source.

Figure 4.11 depicts the difference wave that was obtained by subtracting the trial average for standard tones from the trial average for deviant tones. A negative deflection after 100 ms is clearly visible. The event-related field indicates patterns of activity at central channels in both hemispheres. These findings are consistent with the mismatch negativity (Garrido et al., 2009). We now proceed to localizing the sources of the activation induced by the mismatch negativity.

Figure 4.13 depicts the localized sources. The spatial prior leads to stronger source currents which are spread over a somewhat larger brain volume. The model has correctly identified the source over left temporal cortex, but it does not capture the source over right temporal cortex that should also be present (cf. Fig. 4.11). Note however, that the source estimates in Fig. 4.13 represent estimated mean power, that is, the approximate posterior mean of the $\boldsymbol{\beta}$ parameters. Differences between the decoupled and the coupled prior become more salient when we look at the approximate posterior variances of the auxiliary variables shown in Fig. 4.13. The panels show that sources in both left and right hemispheres are relevant. The relevance of the source in the right hemisphere becomes more pronounced when using the coupled prior.

Discussion

The results in the previous section show that for a real-world data-set, the coupled spatial prior yields importance maps that provide more information for source localization than the approximate posterior mean $\mathrm{E}_q[\boldsymbol{\beta}]$ of the regression coefficients. The visualization of the variance of the auxiliary variables gives additional insight into the relevance of the source currents in the context of a mismatch negativity experiment.

4.5.2 Multivariate fMRI analysis

We applied the logistic regression model in a multivariate fMRI analysis setting to detect areas of brain activation related to an image classification task. We used the blood oxygenation level dependent (BOLD) signal to record functional images that were further pre-processed and used for defining the data in the logistic regression model. Our goal was to create task-related smooth importance maps from the (approximate) posterior moments of the auxiliary variables.

Problem description

We presented to the subject the images of sixes and nines from the MNIST[4] data-set and we recorded the BOLD response (Penny et al., 2005). The variables $y_i \in \{-1, 1\}$ denote the class of the image and the BOLD response was pre-processed to form the rows $\boldsymbol{x}_i$ of the design matrix $\boldsymbol{X} \in \mathbb{R}^{m,n}$. We assumed that the data $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, m$ are independent and identically distributed. Our aim is to use the approximate posterior marginal variances of the components of $\boldsymbol{u}$ to create smooth importance maps and to investigate whether the coupled prior yields better performances than the uncoupled one. In this setting, we also consider spatio-temporal modeling, that is, the vector of coefficients $\boldsymbol{\beta}$ consists of the concatenation of the spatial coefficients $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(T)}$ for time steps $t = 1, \ldots, T$. The corresponding variables in $\boldsymbol{u}^{(t)}$ and $\boldsymbol{v}^{(t)}$, $t = 1, \ldots, T$ are also *temporally* coupled through $\boldsymbol{Q}$. Details follow in the sections below. Figure 4.4 shows a few re-scaled handwritten digits used in the experiment.

Experimental setting

The data was collected for one subject and the experiment consisted of the following. First, we interpolated the 28×28 pixel gray-scale images of the MNIST data-set to 256×256 pixel images. In each trial, the image of a six or nine was presented to the subject for 12.5 seconds. The stimuli images flickered at a 6 Hz rate on a black background. The task of the subject was to maintain focus on a fixation dot and to detect and signal a brief change in color of the fixation dot from red to green and back to red. The duration of the color change was 0.03 seconds and it happened once per trial at random times. The subject was asked to signal the detection of the color transition by pressing a button. One hundred trials were performed with an inter-trial interval of 12.5 seconds.

The functional images were recorded using a *Siemens 3 T* MRI system with a 32 channel coil for signal reception. The BOLD functional images were acquired using a

[4]The handwritten digit image data-set is available at http://yann.lecun.com/exdb/mnist.

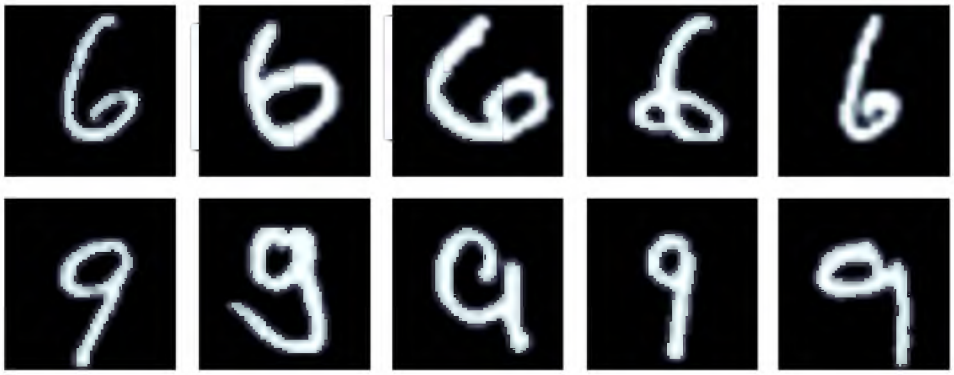

Figure 4.4: Example of the variations within the set of handwritten sixes and nines.

single shot gradient EPI sequence, with a repetition time of $TR = 2.5$ seconds (this means a total of $T = 6$ time-slices), an echo time of $TE = 0.03$ seconds and an isotropic voxel size of $2 \times 2 \times 2$ mm. There were 42 axial slices recorded in an ascending order. A high resolution anatomical image was acquired using an MP-RAGE sequence with $TE/TR = 3.39/2500$, 176 sagittal slices and an isotropic voxel size of $1 \times 1 \times 1$ mm.

The functional images were pre-processed within the framework of SPM5[5] (Statistical Parametric Mapping). In order to correct for motion, the functional volumes were realigned to the mean image. The anatomical image was co-registered with the mean of the functional images. The number of voxels was reduced by applying a gray-matter mask with threshold $p > 0.5$ and a voxel size of $4 \times 4 \times 4$ mm. The functional data was high-pass filtered and de-trended. The volume information acquired up to 15 seconds after the beginning of the trial was collected to estimate the response in individual voxels. The trial data was normalized, that is, the response data were re-scaled to have zero mean and unit variance.

Specifying the prior

In addition to the prior defined in Section 4.5.1, which we will call *spatial prior*, we introduce a *spatio-temporal prior*. In this prior, besides coupling the spatial components, we introduce a temporal coupling of the variables, that is, we add terms of the form $s_{\text{temp}}(u_k^{(t+1)} - u_k^{(t)})^2$, $t = 1, \ldots, T-1$ to the log-prior. This leads to a prior that is similar to

$$p(\boldsymbol{u}) \propto \exp\Big\{ -\frac{1}{2}\sum_t \sum_k (u_k^{(t)})^2 - \frac{1}{2} s_{\text{temp}} \sum_t \sum_k (u_k^{(t+1)} - u_k^{(t)})^2 \ldots \\ - \frac{1}{2} s_{\text{spt}} \sum_t \sum_{i \sim j} (u_i^{(t)} - u_j^{(t)})^2 \Big\}, \quad (4.10)$$

[5] Available at http://fil.ion.ucl.ac.uk/spm.

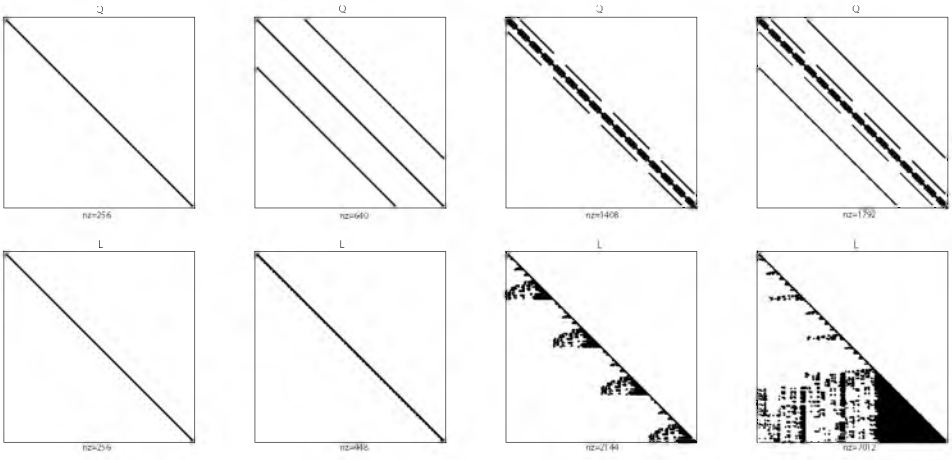

Figure 4.5: Sparsity pattern of precision matrices (top row) and lower-triangular matrices $\boldsymbol{L}$ (bottom row) after a Cholesky decomposition and reordering using the AMD algorithm for a $4\times4\times4\times4$ volume. From left to right, a sparse prior, temporal prior, spatial prior, and spatio-temporal prior were used. The number of non-zero elements are shown below each matrix.

but parameterized as described in Section 4.5.1. The precision matrix $\boldsymbol{Q}$ has three parameters: s_{spt}, s_{temp} and θ. In the experiments, we use only one coupling parameter $s = s_{\text{spt}} = s_{\text{temp}}$.

Figure 4.5 illustrates the structure of the precision matrix $\boldsymbol{Q}$ and its AMD reordered Cholesky factor $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{Q}$ for the various problem settings. For illustration purposes, we considered a toy $4 \times 4 \times 4$ grid and $T = 4$ time-steps. Note that, as to be expected, the spatio-temporal prior has the highest number of non-zeros both in $\boldsymbol{Q}$ and $\boldsymbol{L}$ and compared to the spatial coupling, the spatio-temporal coupling introduces a considerable amount of non-zeros. In order to get an empirical estimate of computation time, we ran the EP algorithm on $M \times M \times M \times M$ volumes with M ranging from one to ten. We used 40 trials per condition and filled the volumes with random voxels from the original fMRI dataset. Figure 4.6 shows the number of non-zeros and computation time for the different priors as we vary the volume dimensions. Even though we have an exponential increase in the number of non-zero elements and computation time with dimensionality, we are still able to handle very large models with the less complex priors. Note that the increase in computation time for the more complex spatio-temporal models is not mainly due to the increased number of non-zero elements in $\boldsymbol{Q}$ itself, but rather to the disproportionate increase in the number of non-zero elements it implies in $\boldsymbol{L}$.

Results

The spatial model

The experimental data for the spatial model was created by averaging the last three measurements $t = 4, 5, 6$ in every trial. Figure 4.7 shows that the predictive performance

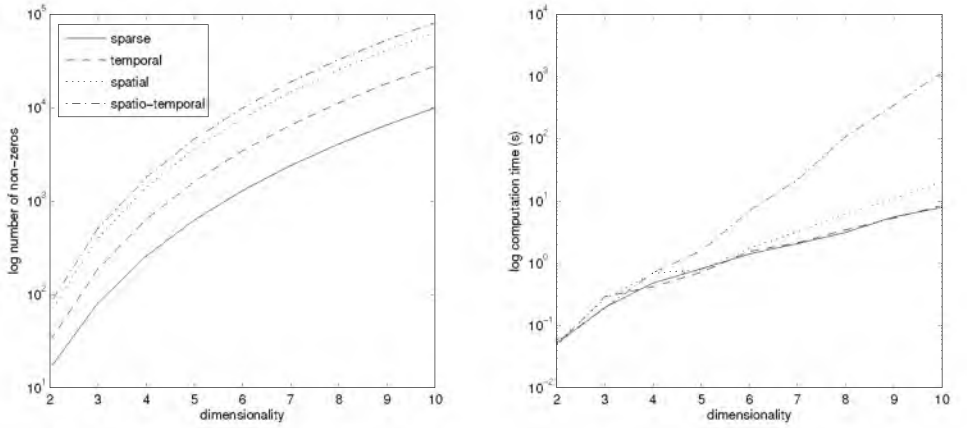

Figure 4.6: Number of non-zero elements in $\boldsymbol{Q}$ and computation time of the EP algorithm as a function of the dimensionality M of the volume when using a spatial prior. (The experiments were carried out on a machine with a 64 bit Intel Xeon 2.83 GHz CPU and 16 GB internal memory.)

of the spatial model is marginally better than that of the decoupled model, although the difference is not significant. Furthermore, the results clearly show that optimal performance is reached when $\theta \approx 0.01$, indicating the improvement over the non-regularized model that is obtained in the limit when θ goes to infinity. Likewise, too much regularization is also detrimental to predictive performance. Predictive performance was significant ($p < 0.05$) for all models between $\theta = 0.01$ and $\theta = 1$. Log model evidence approximation was slightly larger for the spatial model ($\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta = 0.01, s = 10) \approx -3193.84$) as compared with the decoupled model ($\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta = 0.01, s = 0) \approx -3194.21$).

We emphasized the improved interpretability that is obtained when using informed priors. Figure 4.8 establishes this claim by showing the number of included voxels sorted according to importance versus the number of clusters obtained, where a cluster is defined as a connected component in the measured brain volume. The spatial prior leads to a much lower number of clusters compared to the decoupled prior. The absolute number of clusters remains relatively large due to the gray-matter mask, which selects a non-contiguous subset of voxels from the measured volume.

Figure 4.12 provides a visualization of the resulting models. The spatial prior leads to the selection of clusters of important voxels. Spatial regularization can also be interpreted as a form of noise reduction since spatially segregated voxels are less likely to have large importance values. The mean regression coefficients have a large magnitude for the most important voxels. Note the strong agreement between the uncoupled and the spatial model regarding these voxels. It can also be seen from Fig 4.12 that in the uncoupled model the most important voxels are to some extent scattered throughout the brain, while in the spatially coupled model, they are almost exclusively observed in the occipital lobe encompassing Brodmann Areas 17 and 18. This pattern of results appears neuro-biologically plausible, the only difference between conditions being visual.

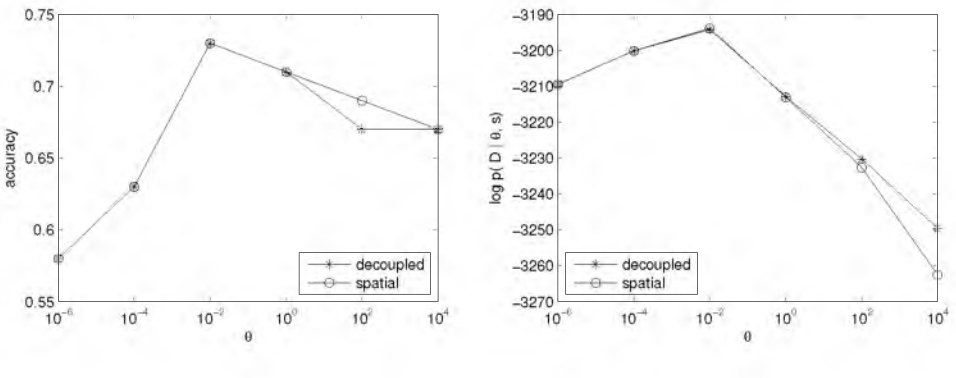

Figure 4.7: Proportion of correctly classified trials and approximate log model evidence for the decoupled model and the spatial model.

The spatio-temporal model

Figure 4.9 shows that the predictive performance of the temporal model is again somewhat better than that of the decoupled model although the difference is not significant. Optimal performance is reached when $\theta = 1$. Predictive performance was only significant ($p < 0.05$) for this setting of the scale parameter. Log model evidence was slightly larger for the decoupled model ($\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta = 10^{-4}, s = 0) \approx -13629.79$) as compared with the temporal model ($\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta = 10^{-4}, s = 10) \approx -13629.93$). Note the disagreement between the optimum according to predictive performance and according to model evidence; it is well-known that model evidence optimization does not always lead to the best predicting models. Note further that predictive performance of the temporal model was significantly lower than that of the spatial model due to the inclusion of volumes for which the task-related response is likely to be negligible.

Figure 4.10 depicts the importance values for five consecutive volumes for ten voxels that were considered to be most important by the spatial model. The temporal model leads to temporal smoothing of the importance values. As a result, it becomes clear that the last few volumes carry more task-related information, which is in agreement with the (lagged) BOLD response.

Discussion

The results show that the sparsity inducing multivariate prior we defined can be successfully applied in the context of multivariate fMRI analysis. Although it does not have a very strong impact on the prediction performance of the model, the coupled prior can provide smooth importance maps that are more biologically plausible as the relevant variables are not scattered throughout the whole brain volume. The majority of the relevant coefficients, however, show a similar spatial pattern as for the uncoupled case. This allows us to conclude that the use of the coupled prior induces the smoothing effect that is desirable in many applications where the localization of the active brain areas is important.

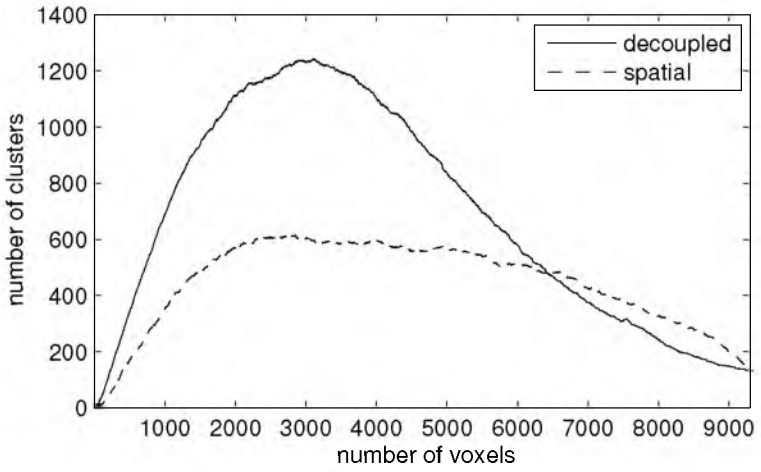

Figure 4.8: Number of clusters obtained when varying the number of included voxels sorted according to importance for the decoupled model and the spatial model.

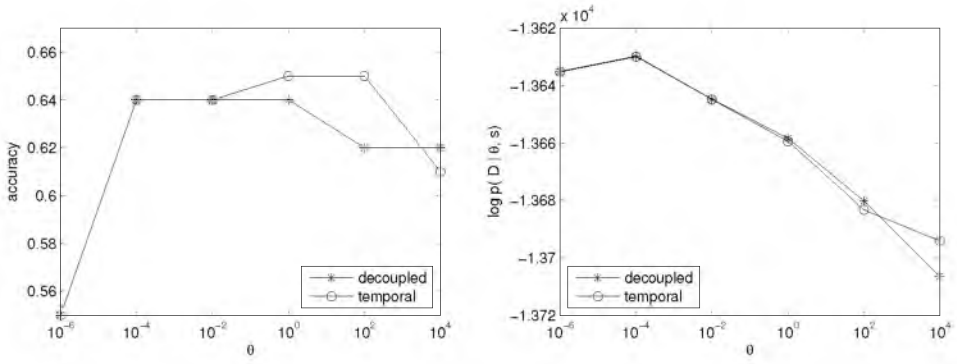

Figure 4.9: Proportion of correctly classified trials and approximate log model evidence for the decoupled model and the temporal model.

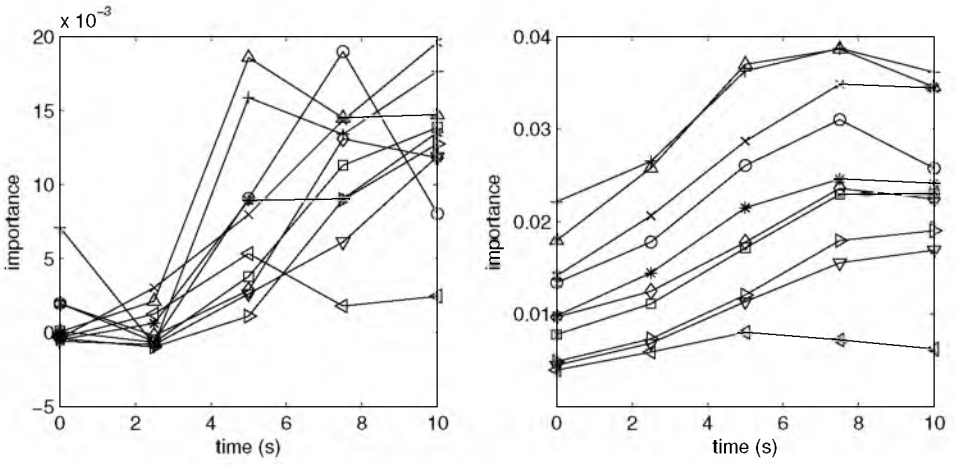

Figure 4.10: Importance values for ten voxels whose BOLD response was acquired over five consecutive volumes using the decoupled model (left) and the temporal model (right).

4.6 Conclusions

In this chapter, we introduced a multivariate sparsity inducing scale mixture distribution and we have shown that it can be applied to cognitive neuroscience problems. This distribution can be viewed as a multivariate generalization of the double exponential distribution and it is somewhat similar in spirit to the scale mixture distribution introduced by Lyu and Simoncelli (2006) for analyzing photographic images. The motivation behind our approach was to introduce prior correlations between the magnitudes of the regression coefficients in order to represent the inherent spatial and spatio-temporal smoothness in task-related brain activity. The introduced hierarchical prior keeps the regression coefficients β_j uncorrelated, but it couples their magnitudes. By this, it avoids the drawback of the multivariate double exponential prior in Eltoft et al. (2006) which can be viewed as a prior acting on $\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}$. The symmetry in the auxiliary variables $\boldsymbol{u}$ and $\boldsymbol{v}$ implies that the posterior density leads to a block-diagonal covariance structure and that approximate inference with EP can be performed very efficiently with complexities that scale roughly linearly in terms of the number of variables n.

Our choice for a double exponential distribution related multivariate prior was motivated by the success of the independent double exponential priors both in the MAP (e.g. Tibshirani, 1996; Cawley et al., 2007; Williams, 1995) and the Bayesian setting (Seeger, 2008). The experiments on real-world data show that the scale mixture distribution we introduced can take into account spatial an spatio-temporal smoothness properties and lead to meaningful results and fast approximate inference algorithms.

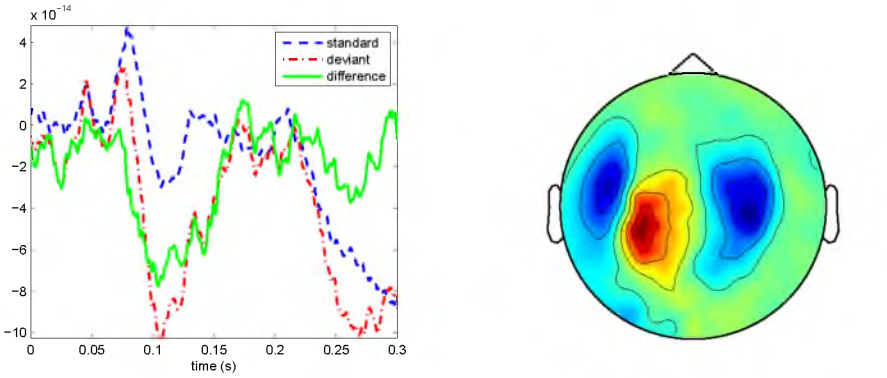

Figure 4.11: Evolution of the difference wave at right central sensors and event-related field of the difference wave 125 ms after cue onset.

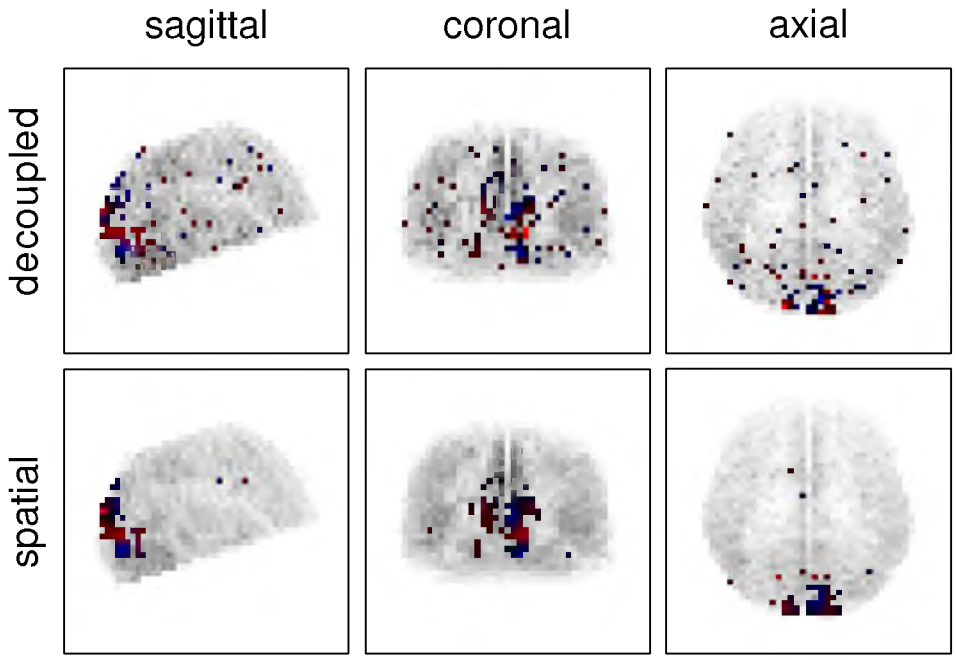

Figure 4.12: Glass-brain view of importance values for the decoupled versus the spatial prior. Means of the regression coefficients of the 100 most important voxels are color-coded with red standing for positive and blue for negative values.

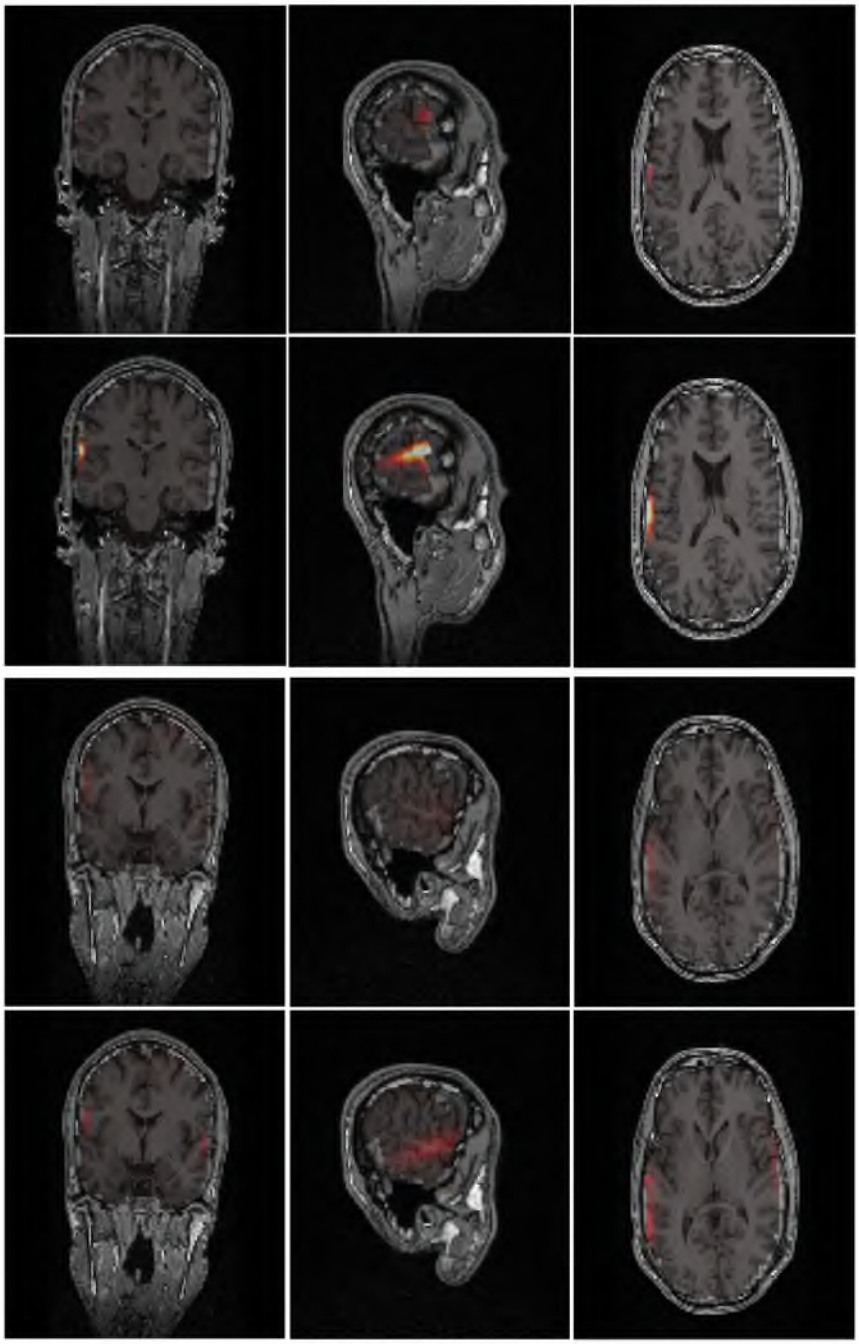

Figure 4.13: Source estimates using a decoupled prior (1st row) or a coupled prior (2nd row). The variance of the auxiliary variables using a decoupled prior (3rd row) or a coupled prior (4th row).

Appendix A

A.1 Properties and proofs

Lemma A1. (Watanabe and Fukumizu, 2009) *For any graph $G = (V, E)$, edge adjacency matrix $\boldsymbol{\mathcal{M}}(\alpha)$ (defined in Section 2.4.1), and arbitrary vector $\boldsymbol{w} \in \mathbb{R}^{|E|}$ and the corresponding graph one has*

$$\det\left(\boldsymbol{I}_{|E|} - \alpha^{-1}\mathrm{diag}\left(\boldsymbol{w}\right)\boldsymbol{\mathcal{M}}(\alpha)\right) = \det\left(\boldsymbol{I}_{|V|} + \alpha^{-1}\boldsymbol{A}(\boldsymbol{w})\right)\prod_{ij}(1 - w_{ij}w_{ji}),$$

where

$$A_{ii}\left(\boldsymbol{w}\right) = \sum_{i\sim j}\frac{w_{ij}w_{ji}}{1 - w_{ij}w_{ji}} \qquad \textit{and} \qquad A_{ij}\left(\boldsymbol{w}\right) = -\frac{w_{ij}}{1 - w_{ij}w_{ji}}.$$

Proof: In the following, we reproduce their proof in a somewhat simpler form. Let us define $\boldsymbol{U}_{ij,\cdot} = \boldsymbol{e}_j^T$, $\boldsymbol{V}_{ij,\cdot} = \boldsymbol{e}_i^T$—where $\boldsymbol{e}_k$ is the k^{th} unit vector of $\mathbb{R}^n$—and $\boldsymbol{S}$ with

$$\begin{bmatrix} S_{ij,ij} & S_{ij,ji} \\ S_{ji,ij} & S_{ji,ji} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then we have $\boldsymbol{\mathcal{M}}(\alpha) = \boldsymbol{U}\boldsymbol{V}^T - \alpha\boldsymbol{S}$. Let us define $\boldsymbol{W} \in \mathbb{R}^{|E|\times|E|}$ a diagonal matrix with $w_{ij,ij} = w_{ij}$. Using the matrix determinant lemma this reads as

$$\begin{aligned}
&\det\left(\boldsymbol{I} - \alpha^{-1}\boldsymbol{W}\left(\boldsymbol{U}\boldsymbol{V}^T - \alpha\boldsymbol{S}\right)\right) \\
&\quad = \det\left(\boldsymbol{I} + \boldsymbol{W}\boldsymbol{S} - \alpha^{-1}\boldsymbol{W}\left(\boldsymbol{U}\boldsymbol{V}^T\right)\right) \\
&\quad = \det\left(\boldsymbol{I} - \alpha^{-1}\boldsymbol{W}\left(\boldsymbol{U}\boldsymbol{V}^T\right)\left(\boldsymbol{I} + \boldsymbol{W}\boldsymbol{S}\right)^{-1}\right)\det\left(\boldsymbol{I} + \boldsymbol{W}\boldsymbol{S}\right) \\
&\quad = \det\left(\boldsymbol{I} - \alpha^{-1}\boldsymbol{V}^T\left(\boldsymbol{I} + \boldsymbol{W}\boldsymbol{S}\right)^{-1}\boldsymbol{W}\boldsymbol{U}\right)\det\left(\boldsymbol{I} + \boldsymbol{W}\boldsymbol{S}\right).
\end{aligned}$$

The (ij, ji) block of $(\boldsymbol{I} + \boldsymbol{WS})^{-1}\boldsymbol{W}$ is

$$\frac{1}{1-w_{ji}w_{ji}}\begin{bmatrix} 1 & -w_{ij} \\ -w_{ji} & 1 \end{bmatrix}\begin{bmatrix} w_{ij} & 0 \\ 0 & w_{ji} \end{bmatrix} = \frac{1}{1-w_{ji}w_{ji}}\begin{bmatrix} w_{ij} & -w_{ij}w_{ji} \\ -w_{ji}w_{ij} & w_{ji} \end{bmatrix}$$

and thus, we can define $\boldsymbol{A} \equiv \boldsymbol{V}^T(\boldsymbol{I} + \boldsymbol{WS})^{-1}\boldsymbol{WU}$ such that

$$A_{i,i} = \sum_{i\sim j}\frac{w_{ij}w_{ji}}{1-w_{ij}w_{ji}} \quad \text{and} \quad A_{i,j} = -\frac{w_{ij}}{1-w_{ij}w_{ji}}.$$

This completes the proof of the matrix determinant lemma (2.25) in Section 2.4.2. ■

Property A1. *The matrix $\boldsymbol{\mathcal{M}}(\alpha) = \boldsymbol{UV}^T - \alpha\boldsymbol{S}$ is singular only for K-regular graphs with $\alpha = K$.*
Proof: Let $x \in \mathbb{R}^{|E|}$ and $\boldsymbol{y} = \boldsymbol{\mathcal{M}}(\alpha)\boldsymbol{x}$. Then $y_{ij} = \sum_{k\sim j} x_{jk} - \alpha x_{ji}$. Let us fix j, then $y_{ij} = 0$ for any i means that $\sum_{k\sim j} x_{jk} = \alpha x_{ji}$ for any i. This can only hold if the graph is K-regular, $\alpha = K$ and all x_{ij}s are equal or $x_{ij} = 0$ for all pairs ij. ■

Property A2. *For a suitable chosen $\epsilon > 0$ there always exists an α_ϵ such that the constrained fractional free energy F^c_α possesses a local minimum for all $0 < \alpha < \alpha_\epsilon$.*
Proof: Let us define $\boldsymbol{v}^*_{MF} = \operatorname{argmin}_{\boldsymbol{v}} F_{MF}(\boldsymbol{v})$ and

$$U^\epsilon_{MF} = \{\boldsymbol{v} : F_{MF}(\boldsymbol{v}) \leq F_{MF}(\boldsymbol{v}^*_{MF}) + 2\epsilon\}.$$

The form of F_{MF} implies that we can always choose ϵ such that U^ϵ_{MF} is a proper subset of the positive "quadrant" in $\mathbb{R}^n$, in other words, $U^\epsilon_{MF} \subset \mathbb{R}^n_+$. Then due to the properties of F_{MF} (continuous and convex, with a unique finite global minimum attained at a finite value), the domain U^ϵ_{MF} is closed, bounded, convex and $\boldsymbol{v}^*_{MF} \in U^\epsilon_{MF} \setminus \partial U^\epsilon_{MF}$, that is, $\boldsymbol{v}^*_{MF}$ is in the interior of U^ϵ_{MF}. Since F_{MF} and $F^c_\alpha(\boldsymbol{v})$ are continuous on $\mathbb{R}^n_+$, the set U^ϵ_{MF} is closed and bounded and $\lim_{\alpha\to 0} F^c_\alpha(\boldsymbol{v}) = F_{MF}(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{R}^n_+$, it follows that F^c_α is converges uniformly on U^ϵ_{MF} as $\alpha \to 0$. This implies that there exists α_ϵ such that $F_{MF}(\boldsymbol{v}_{MF}) - \epsilon < F^c_\alpha(\boldsymbol{v}_{MF}) < F_{MF}(\boldsymbol{v}_{MF})$ for all $0 < \alpha < \alpha_\epsilon$ and all $\boldsymbol{v} \in U^\epsilon_{MF}$. Let us fix α. It is known that, since U^ϵ_{MF} is closed and bounded and F^c_α is continuous, F^c_α attains its extrema on U^ϵ_{MF}. Since $F_{MF}(\boldsymbol{v}) = F_{MF}(\boldsymbol{v}^*_{MF}) + 2\epsilon$ for all $\boldsymbol{v} \in \partial U^\epsilon_{MF}$ and $F^c_\alpha(\boldsymbol{v}) > F_{MF}(\boldsymbol{v}) - \epsilon$ for all $\boldsymbol{v} \in U^\epsilon_{MF}$ it follows that $F^c_\alpha(\boldsymbol{v}) > F_{MF}(\boldsymbol{v}^*_{MF}) + \epsilon$ for all $\boldsymbol{v} \in \partial U^\epsilon_{MF}$. We have chosen α such that $F_{MF}(\boldsymbol{v}^*_{MF}) - \epsilon < F^c_\alpha(\boldsymbol{v}^*_{MF}) < F_{MF}(\boldsymbol{v}^*_{MF})$. The latter two conditions imply that one of the extrema has to be a local minimum in the interior of U^ϵ_{MF}. ■

A.2 Solving the Takahashi equations

The Takahashi equations (Takahashi et al., 1973) aim to compute certain elements of the inverse of a positive definite matrix from its Cholesky factor. The derivation of the equations or the algorithm can be found in many papers (e.g. Erisman and Tinney, 1975). In the following we present the line of argument in Rue et al. (2009). Let $\boldsymbol{Q} = \boldsymbol{LL}^T$, $\boldsymbol{z} \sim N(0, \boldsymbol{I})$ and $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$. Then using the notation $\boldsymbol{V} = \boldsymbol{Q}^{-1}$ we find that $\boldsymbol{x} \sim N(0, \boldsymbol{V})$.

The equation $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$ leads to $L_{ii}x_i = z_i - L_{ii}^{-1}\sum_{k=i+1}^{n} L_{ki}x_k$. Multiplying both sides with $x_j, j \geq N$, using $\boldsymbol{z} = \boldsymbol{L}^{-T}\boldsymbol{x}$ and taking expectations we arrive at the Takahashi equations $V_{ij} = \delta_{ij}L_{ii}^{-2} - L_{ii}^{-1}\sum_{k=i+1}^{n} L_{ki}V_{kj}$. Since we only want to compute the diagonal of $\boldsymbol{V}$ or the elements V_{ij} for which $L_{ij} \neq 0$, the algorithm can be written in the following MATLAB friendly form

1: **function** $\boldsymbol{V}$ = `SolveTakahashi`($\boldsymbol{L}$)
2: **for** $i = n : -1 : 1$
3: $I = \{j : L_{ij} \neq 0,\ j > i\}$
4: $\boldsymbol{V}_{I,i} = -\boldsymbol{V}_{I,I}\boldsymbol{L}_{I,i}/L_{i,i}$
5: $\boldsymbol{V}_{i,I} = \boldsymbol{V}_{I,i}^T$
6: $V_{i,i} = 1/L_{i,i}^2 - \boldsymbol{V}_{i,I}\boldsymbol{L}_{I,i}/L_{ii}$
7: **end**

The complexity of this algorithm scales roughly with $nnzeros(\boldsymbol{Q})^2/n$.

A.3 Gaussian formulas

The first and second moments of a distribution $p(\boldsymbol{x}) = Z^{-1}(\boldsymbol{m}, \boldsymbol{V})f(\boldsymbol{x})\,N(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{V})$ are given by

$$\begin{aligned} \mathrm{E}_p[\boldsymbol{x}] &= \boldsymbol{m} + \boldsymbol{V}\nabla_{\boldsymbol{m}} \log Z(\boldsymbol{m}, \boldsymbol{V}), \\ \mathrm{V}_p[\boldsymbol{x}] &= \boldsymbol{V} + \boldsymbol{V}\nabla^2_{\boldsymbol{mm}} \log Z(\boldsymbol{m}, \boldsymbol{V})\boldsymbol{V}. \end{aligned} \tag{A.1}$$

Applying integration by parts, one can show that the moments of p can also be written in the form

$$\begin{aligned} \mathrm{E}_p[\boldsymbol{x}] &= \boldsymbol{m} + \frac{1}{Z}\boldsymbol{V}\mathrm{E}_q[\nabla_{\boldsymbol{x}} f], \\ \mathrm{V}_p[\boldsymbol{x}] &= \boldsymbol{V} + \frac{1}{Z^2}\boldsymbol{V}\left[Z\mathrm{E}_q\left[\nabla^2_{\boldsymbol{xx}} f\right] - \mathrm{E}_q[\nabla_{\boldsymbol{x}} f]\,\mathrm{E}_q[\nabla_{\boldsymbol{x}} f]^T\right]\boldsymbol{V}, \end{aligned} \tag{A.2}$$

provided that $f(\boldsymbol{x})\,e^{-\boldsymbol{x}^T\boldsymbol{x}}$ and $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}e^{-\boldsymbol{x}^T\boldsymbol{x}}$ vanish at infinity and the required differentials and integrals exist.

A.4 Details of EP in latent Gaussian models

Assume the distribution to be approximated has the form

$$p(\boldsymbol{x}) \propto p_0(\boldsymbol{x}) \prod_i t_i(\boldsymbol{U}_i\boldsymbol{x}),$$

where the $\boldsymbol{U}_i$s are matrices that transform the variables into some typically lower dimensional spaces. This formulation includes both the representations when t_j depend only on a subset of parameters, that is, $t_i(\boldsymbol{x}) = t_i(\boldsymbol{x}_{I_i})$ with $\boldsymbol{U}_i = \boldsymbol{I}_{,I_i}$ and the representation used in logistic regression, where $\boldsymbol{U}_i$ is the i^{th} row of the design matrix. Here we present the details of the α-fractional or power EP, where the updates are performed on $\tilde{t}_i^{\alpha}(\boldsymbol{x})$, $\alpha \in (0, 1]$.

Computing $\tilde{t}_i^{new}$

First we compute the form of the term approximations, and show that $\tilde{t}_i$ has a low rank representation. Let $q(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{V})$ and let $\tilde{\boldsymbol{h}} = \boldsymbol{V}^{-1}\boldsymbol{m}$, $\tilde{\boldsymbol{Q}} = \boldsymbol{V}^{-1}$ the canonical parameters of $q(\boldsymbol{x})$. We use $q^{\setminus i}(\boldsymbol{x}) = N\left(\boldsymbol{x}|\boldsymbol{m}^{\setminus i}, \boldsymbol{V}^{\setminus i}\right)$ to denote the distribution $q^{\setminus i}(\boldsymbol{x}) \propto q(\boldsymbol{x})/\tilde{t}_i^{\alpha}(\boldsymbol{x})$. After some calculus, one can show that the moment matching Gaussian $q^{new}(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{m}^{new}, \boldsymbol{V}^{new})$ of $q_i(\boldsymbol{x}) \propto t_i^{\alpha}(\boldsymbol{x}) q^{\setminus i}(\boldsymbol{x})$ is given by

$$\begin{aligned}
\boldsymbol{m}^{new} &= \boldsymbol{m}^{\setminus i} + \boldsymbol{V}^{\setminus i}\boldsymbol{U}_i^T \left[\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right]^{-1} \left[\mathrm{E}[\boldsymbol{z}_i] - \boldsymbol{U}_i \boldsymbol{m}^{\setminus i}\right], \\
\boldsymbol{V}^{new} &= \boldsymbol{V}^{\setminus i} + \boldsymbol{V}^{\setminus i}\boldsymbol{U}_i^T \left[\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right]^{-1} \left[\mathrm{V}[\boldsymbol{z}_i] - \boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right] \left[\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right]^{-1} \boldsymbol{U}_i \boldsymbol{V}^{\setminus i},
\end{aligned}$$

where $\boldsymbol{z}_i$ is a random variable distributed as $\boldsymbol{z}_i \sim t(\boldsymbol{z}_i)^{\alpha} N\left(\boldsymbol{z}_i | \boldsymbol{U}_i \boldsymbol{m}^{\setminus i}, \boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right)$. The update for the term approximation $\tilde{t}_i(\boldsymbol{x})$ is given by $\left(\tilde{t}_i^{new}(\boldsymbol{x})\right)^{\alpha} \propto q^{new}(\boldsymbol{x})/q^{\setminus i}(\boldsymbol{x})$. The latter division yields

$$[\boldsymbol{V}^{new}]^{-1} - \left[\boldsymbol{V}^{\setminus i}\right]^{-1} = \boldsymbol{U}_i^T \left[\mathrm{V}[\boldsymbol{z}_i]^{-1} - \left[\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right]^{-1}\right] \boldsymbol{U}_i \tag{A.3}$$

$$[\boldsymbol{V}^{new}]^{-1} \boldsymbol{m}^{new} - \left[\boldsymbol{V}^{\setminus i}\right]^{-1} \boldsymbol{m}^{\setminus i} = \boldsymbol{U}_i^T \left[\mathrm{V}[\boldsymbol{z}_i]^{-1} \mathrm{E}[\boldsymbol{z}_i] - \left[\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T\right]^{-1} \boldsymbol{U}_i \boldsymbol{m}^{\setminus i}\right] \tag{A.4}$$

leading to

$$\tilde{t}_i^{new}(\boldsymbol{x}) \propto \exp\left((\boldsymbol{U}_j \boldsymbol{x})^T \tilde{\boldsymbol{h}}^j - \frac{1}{2} (\boldsymbol{U}_j \boldsymbol{x})^T \tilde{\boldsymbol{Q}}^j (\boldsymbol{U}_j \boldsymbol{x}) \right),$$

where $\tilde{\boldsymbol{h}}^i$ and $\tilde{\boldsymbol{Q}}^i$ are given by the corresponding quantities in (A.3) and (A.4). The approximating distribution q is defined by the canonical parameters

$$\begin{aligned}
\tilde{\boldsymbol{h}} &= \boldsymbol{h} + \sum_i \boldsymbol{U}_i^T \tilde{\boldsymbol{h}}^i \\
\tilde{\boldsymbol{Q}} &= \boldsymbol{Q} + \sum_i \boldsymbol{U}_i^T \tilde{\boldsymbol{Q}}^i \boldsymbol{U}_i,
\end{aligned}$$

that is, the sum over the parameters of $\tilde{t}_i$ and the parameters of the prior $p_0(\boldsymbol{x}) \propto \exp\left(\boldsymbol{h}^T \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}/2\right)$.

Computing the cavity distribution $q^{\setminus i}$

Now, we turn our attention to the computation of the distribution $q^{\setminus i}$. The quantities we are interested in are $\boldsymbol{U}_i \boldsymbol{m}^{\setminus i}$ and $\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_j^T$. After some calculus, one can show that these are given by

$$\begin{aligned}
\boldsymbol{U}_i \boldsymbol{V}^{\setminus i} \boldsymbol{U}_i^T &= \boldsymbol{U}_i \left(\tilde{\boldsymbol{Q}} - \alpha \boldsymbol{U}_i^T \tilde{\boldsymbol{Q}}^i \boldsymbol{U}_i\right)^{-1} \boldsymbol{U}_i^T \\
&= \left(\boldsymbol{U}_i \boldsymbol{V} \boldsymbol{U}_i^T\right) \left(\boldsymbol{I} - \alpha \tilde{\boldsymbol{Q}}^i \left(\boldsymbol{U}_i \boldsymbol{V} \boldsymbol{U}_i^T\right)\right)^{-1}
\end{aligned}$$

$$\begin{aligned} \boldsymbol{U}_i \boldsymbol{m}^{\backslash i} &= \boldsymbol{U}_i \left(\tilde{\boldsymbol{Q}} - \alpha \boldsymbol{U}_i^T \tilde{\boldsymbol{Q}}^i \boldsymbol{U}_i \right)^{-1} \left(\tilde{\boldsymbol{h}} - \alpha \boldsymbol{U}_i^T \tilde{\boldsymbol{h}}^i \right) \\ &= \left(\boldsymbol{I} - \alpha \tilde{\boldsymbol{Q}}^i \left(\boldsymbol{U}_i \boldsymbol{V} \boldsymbol{U}_i^T \right) \right)^{-1} \left(\boldsymbol{U}_i \boldsymbol{m} - \alpha \left(\boldsymbol{U}_i \boldsymbol{V} \boldsymbol{U}_i^T \right) \tilde{\boldsymbol{h}}^i \right). \end{aligned}$$

Therefore, the computational bottleneck of EP reduces to the computation of the quantities $\boldsymbol{U}_i \boldsymbol{m}$ and $\boldsymbol{U}_i \boldsymbol{V} \boldsymbol{U}_i^T$. These can be computed from the canonical representation of q by $\boldsymbol{U}_i \tilde{\boldsymbol{Q}}^{-1} \boldsymbol{h}$ and $\boldsymbol{U}_i \tilde{\boldsymbol{Q}}^{-1} \boldsymbol{U}_i^T$.

Computing the EP's evidence approximation

Let us define

$$\log Z \left(\boldsymbol{m}, \boldsymbol{V} \right) \equiv \frac{1}{2} \boldsymbol{m}^T \boldsymbol{V}^{-1} \boldsymbol{m} + \frac{1}{2} \log \det \boldsymbol{V} + \frac{n}{2} \log \left(2\pi \right)$$

and

$$\log Z_i \left(\boldsymbol{m}, \boldsymbol{V} \right) \equiv \log \int d\boldsymbol{x} \, N \left(\boldsymbol{x} | \boldsymbol{m}, \boldsymbol{V} \right) t_i^{\alpha} \left(\boldsymbol{U}_i \boldsymbol{x} \right).$$

Expectation propagation approximates the evidence $p \left(\boldsymbol{y} | \boldsymbol{\theta} \right)$ by $Z_{ep} = Z^{1-n/\alpha} \prod_i Z_i^{\alpha}$. Using the above introduced notation this can be written as

$$\begin{aligned} \log Z_{EP} = & \log Z \left(\boldsymbol{m}, \boldsymbol{V} \right) \\ & + \frac{1}{\alpha} \sum_i \left[\log Z_j \left(\boldsymbol{m}^{\backslash i}, \boldsymbol{V}^{\backslash i} \right) + \log Z \left(\boldsymbol{m}^{\backslash i}, \boldsymbol{V}^{\backslash i} \right) - \log Z \left(\boldsymbol{m}, \boldsymbol{V} \right) \right], \end{aligned} \tag{A.5}$$

which in the case when t_i depends on $\boldsymbol{U}_i \boldsymbol{x}$ boils down to

$$\begin{aligned} \log Z_{EP} = & \log Z \left(\boldsymbol{m}, \boldsymbol{V} \right) + \frac{1}{\alpha} \sum_i \log Z_j \left(\boldsymbol{U}_i \boldsymbol{m}^{\backslash i}, \boldsymbol{U}_i \boldsymbol{V}^{\backslash i} \boldsymbol{U}_i^T \right) \\ & + \frac{1}{\alpha} \sum_i \left[\log Z \left(\boldsymbol{U}_i \boldsymbol{m}^{\backslash i}, \boldsymbol{U}_i \boldsymbol{V}^{\backslash i} \boldsymbol{U}_i^T \right) - \log Z \left(\boldsymbol{U}_i \boldsymbol{m}, \boldsymbol{U}_i \boldsymbol{V} \boldsymbol{U}_i^T \right) \right]. \end{aligned}$$

Bibliography

P. R. Amestoy, T. A. Davis, and Iain S. D. An approximate minimum degree ordering algorithm. *SIAM J. Matrix Anal. Appl.*, 17(4):886–905, October 1996.

D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, (Series B)*, 36(1):99–102, 1974.

D. Bickson. *Gaussian Belief Propagation: Theory and Application.* PhD thesis, The Hebrew University of Jerusalem, 2009.

A. Birlutiu and T. Heskes. Expectation propagation for rating players in sports competitions. In J. N. Kok, J. Koronacki, R. López de Mántaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Proceedings ECML/PKDD*, volume 4702 of *Lecture Notes in Computer Science*, pages 374–381. Springer, 2007.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. Discussion Paper 2008-31, Duke University Department of Statistical Science, 2008.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 73–80, 2009.

R. Casella and E. I. George. Explaining the Gibbs sampler. *The Americal Statistician*, 46 (3):167–174, 1992.

G. Cawley, N. Talbot, and M. Girolami. Sparse multinomial logistic regression via Bayesian L1 regularisation. In Schölkopf. B., J. Platt, and T Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

R. G. Cowell, A.P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer-Verlag New York, Inc., 1999.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.

L. Csató and M. Opper. Sparse representation for Gaussian process models. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, Cambridge, MA, USA, 2001. MIT Press.

B. Cseke and T. Heskes. Bounds on the Bethe free energy for Gaussian networks. In D. A. McAllester and P. Myllymäki, editors, *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 97–104. AUAI Press, 2008.

B. Cseke and T. Heskes. Improving posterior marginal approximations in latent Gaussian models. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 121–128, 2010a.

B. Cseke and T. Heskes. Properties of Bethe free energies and message passing in Gaussian models. *(submitted to JAIR)*, 2010b.

B. Cseke and T. Heskes. Approximate marginals in latent Gaussian models. *(submitted to JMLR)*, 2010c.

P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel. Trueskill through time: Revisiting the history of chess. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 337–344. MIT Press, Cambridge, MA, 2008.

T. Eltoft, T. Kim, and T. Lee. On the multivariate Laplace distribution. *IEEE Signal Process Lett*, 13(5):300–303, 2006.

A. M. Erisman and W. F. Tinney. On computing certain elements of the inverse of a sparse matrix. *Commun. ACM*, 18(3):177–179, 1975. ISSN 0001-0782.

M. I. Garrido, J. M. Kilner, K. E. Stephan, and K. J. Friston. The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, 120(3):453–463, 2009.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Pattern Analysis and Machine Intelligence*, 6(721-741), 1984.

E. I. George and R.E. McCullogh. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

G. H. Golub and C.F. van Loan. *Matrix computations*. The John Hopkins University Press, third edition edition, 1996.

P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, 1998.

S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2):726–738, 2008.

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 359–366, Cambridge, MA, 2003. The MIT Press.

T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P11015, 2005.

Tom Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16:2379–2413, 2004.

R. A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 2005.

S. Ingram. Minimum degree reordering algorithms: A tutorial, 2006. URL `http://www.cs.ubc.ca/~sfingram/cs517_final.pdf`.

T. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, *Advanced mean field methods: theory and practice*, pages 129–160, Cambridge, MA, 2000. The MIT Press.

J. K. Johnson, D. Bickson, and D. Dolev. Fixing convergence of Gaussian belief propagation. *CoRR*, abs/0901.4192, 2009a.

J. K. Johnson, V. Y. Chernyak, and M. Chertkov. Orbit-product representation and correction of Gaussian belief propagation. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 60–68, 2009b.

S. Kotz, T. J. Kozubowski, and K. Podgórski. *The Laplace distribution and generalizations*. Birkhäuser Boston Inc., Boston, MA, 2001. ISBN 0-8176-4166-1. A revisit with applications to communications, economics, engineering, and finance.

F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2), February 2001.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005. ISSN 1533-7928.

S. L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, New York, USA, July 1996. ISBN 0198522193.

Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.

N. T. Longford. Classes of multivariate exponential and multivariate geometric distributions derived from Markov processes. In H. W. Block, A. R. Sampson, and T. H. Savits, editors, *Topics in statistical dependence*, volume 16 of *IMS Lecture Notes Monograph Series*, pages 359–369. Hayward, CA, 1990.

S. Lyu and E.P. Simoncelli. Statistical modeling of images with fields of Gaussian scale mixtures. In *Advances in Neural Information Processing Systems 19*, pages 945–952. MIT Press, 2006.

D. Malioutov, J. Johnson, and A. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *Journal of Machine Learning Research*, 7:2031–2064, October 2006.

S. Martino and H. Rue. Implementing approximate Bayesian inference using integrated nested Laplace approximation: a manual for the INLA program. Technical report, Departament of Mathematical Sciences, NTNU, Norway, 2009.

L. Meier, S. van der Geer, and P. Bühlman. The group lasso for logistic regression. *Journal Of The Royal Statistical Society (Series B)*, 70(1):53–71, 2008.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):1087–1092, 1953.

T. P. Minka. From hidden Markov models to linear dynamical systems. Technical Report 591, 1998.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

T. P. Minka. Power EP. Technical report, Microsoft Research Ltd., Cambridge, UK, MSR-TR-2004-149, October 2004.

T. P. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd., Cambridge, UK, December 2005.

Ciamac Moallemi and Benjamin Van Roy. Consensus propagation. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 899–906. MIT Press, Cambridge, MA, 2006.

K. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, volume 9, pages 467–475, San Francisco, USA, 1999. Morgan Kaufman.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

Y. Nishiyama and S. Watanabe. Accuracy of loopy belief propagation in Gaussian models. *Neural Networks*, 22(4):385 – 394, 2009. ISSN 0893-6080. doi: DOI: 10.1016/j.neunet.2009.01.003.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural Comput.*, 21(3):786–792, 2009.

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

M. Opper, U. Paquet, and O. Winther. Improving on expectation propagation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1241–1248. MIT, Cambridge, MA, US, 2009.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers, San Mateo, CA, 1988.

W. Penny, N. J. Trujillo-Barreto, and K. J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24:350–362, 2005.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, UK, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal Of The Royal Statistical Society (Series B)*, 71(2):319–392, 2009.

P. Rusmevichientong and B. Van Roy. An analysis of belief propagation on the turbo decoding graph with Gaussian densities. *IEEE Transactions on Information Theory*, 47:745–765, 2001.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008. ISSN 1533-7928.

K. Takahashi, J. Fagan, and M.-S. Chin. Formation of a sparse impedance matrix and its application to short circuit study. In *Proceedings of the 8th PICA Conference*, 1973.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

M. van Gerven, B. Cseke, R. Oostenveld, and T. Heskes. Bayesian source localization with the multivariate Laplace prior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909, 2009.

M. van Gerven, B. Cseke, F. de Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *Neuroimage*, 50(1):150–161, March 2010.

M. Wainwright, T. Jaakkola, and A. Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003.

Y. Watanabe and K. Fukumizu. Graph zeta function in the Bethe free energy and loopy belief propagation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2017–2025. The MIT Press, 2009.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

M. Welling and Y. W. Teh. Belief optimization for binary networks: a stable alternative to loopy belief propagation. In J. S. Breese and D. Koller, editors, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 554–561. Morgan Kaufmann Publishers, 2001.

W. Wiegerinck and T. Heskes. Fractional belief propagation. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 438–445, Cambridge, MA, 2003. The MIT Press.

P. M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

D. Wipf and S. Nagarajan. A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966, February 2009.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems 12*, pages 689–695, Cambridge, MA, 2000. The MIT Press.

O. Zoeter and T. Heskes. Change point problems in linear dynamical systems. *Journal of Machine Learning Research*, 6:1999–2026, 2005a.

O. Zoeter and T. Heskes. Gaussian quadrature based expectation propagation. In Z. Ghahramani and R. Cowell, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 445–452. Society for Artificial Intelligence and Statistics, 2005b.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society (Series B)*, 67(2):301–320, 2005.

Samenvatting

Een veelvoorkomende vorm van inferentie in de Bayesiaanse statisiek is het berekenen van marginale dichtheden of kansen op bepaalde variabelen. Het is vaak onhaalbaar om het exacte antwoord uit te rekenen, zodat benaderingen nodig zijn. In de laatste jaren wordt er daarom steeds meer waarde gehecht aan benaderingsmethoden waarbij een optimalisatieprobleem wordt opgelost met behulp van *variationele benaderingen*. In dit proefschrift presenteren we toepassingen van het verwachtingswaarde-aanpassings algoritme ('expectation propagation', EP), een bekende variatiebenadering voor modellen waarbij de a-priori kansdichtheden normaal verdeeld zijn.

In hoofdstuk 1 introduceren we Bayesiaanse netwerken, grafische modellen en variatiebenaderingen om inferentie in deze modellen te kunnen doen.

In hoofdstuk 2 behandelen we modellen waarbij de variabelen na observatie ook normaal verdeeld zijn en bestuderen we de eigenschappen van het berichtjesuitwisselingsalgoritme (een variant van EP) en de bijbehorende Bethe vrije energie. Ondanks dat de vrije energie wat betreft de benaderde marginale dichtheden (als functionele parameters) dezelfde eigenschappen heeft als in het discrete geval, blijkt het gedrag verrassend zodra deze uitgedrukt wordt in een parametrische vorm waarbij het verzekerd is dat de marginale benaderingen consistent zijn. Als de vrije energie wordt uitgedrukt in momentparameters, is dit in het discrete geval een begrensde functie, maar juist onbegrensd als het model normaal verdeeld is. De bekende versoepelingen die in het discrete geval worden toegepast (e.g. Wainwright et al., 2003; Wiegerinck and Heskes, 2003) lijken het tegengestelde effect te bereiken doordat ze leiden tot een convexe doelfunctie met een onbegrensd globaal minimum. We laten zien dat de stabiele vaste punten van een Gaussische berichtjesuitwisselingsalgoritme lokale minima zijn van een normale vrije energie en dat zowel het convergeren van het berichtjesuitwisselingsalgoritme als het bestaan van lokale minima waarschijnlijker is voor versoepelingsparameters die de vrije energie dichter naar het gemiddelde-veld vrije energie op duwen. We geven ook voldoende en noodzakelijke voorwaarden voor het begrensd zijn van normale Bethe vrije energie.

In hoofdstuk 3 gaan we in op het probleem van het benaderen van marginale kansen in modellen waarin de a-priori kansen normaal verdeeld zijn, maar de variabelen na observatie niet normaal verdeeld zijn, waarbij wordt aangenomen dat de geobserveerde variabelen onafhankelijk en identiek verdeeld zijn gegeven de latente normale variabelen. We stellen methoden voor die niet beperkt hoeven te blijven tot deze modellen, maar er wel uitermate geschikt voor zijn. Marginale dichtheden worden in deze modellen normaal gesproken berekend door een niet-normale dichtheid te benaderen met een multivariate normale dichtheid op basis van de Kullback-Leibler divergentie, oftewel door

het verwachtingswaarde-aanpassings algoritme. In hoofdstuk 3 gaan we een stap verder dan deze methoden en leiden we een raamwerk af om benaderingen met behulp van normale dichtheden te verbeteren. Het raamwerk met deze verbeterde dichtheden laat goede resultaten zien in modellen waar EP convergeert, namelijk wanneer deze dichtheden log-concaaf zijn. Hoewel we met deze benaderingen geen schatting of bovengrens van de fout kunnen geven, hebben de benaderingen de prettige eigenschap dat ze geleidelijk verbeterd kunnen worden wanneer hogere nauwkeurigheid nodig is.

In hoofdstuk 4 definiëren we een multivariaat schaalmengsel (scale mixture) distributie die als a-priori kansverdeling wordt gebruikt om coëfficiënten afkomstig uit lineaire en logistische regressie uit te dunnen. Hiermee leiden we een efficiënt verwachtingswaarde-aanpassings algoritme af om benaderingen uit te rekenen in deze modellen. We passen deze modellen in MEG en fMRI experimenten toe om te bepalen welke hersengebieden worden geactiveerd als de proefpersoon een bepaalde taak moet uitvoeren. We definiëren een multivariate a-priori dichtheid gebaseerd op de scale mixture representatie van de univariate dubbele exponentiële dichtheid met als doel om correlaties tussen de a-priori dichtheden van de absolute waarde van de regressiecoëfficiënten te introduceren. Dit werd ondersteund door de observatie dat de activaties in veel MEG- en fMRI-toepassingen geleidelijk veranderende ruimte- en tijdspatronen hebben, dat wil zeggen, naastgelegen hersengebieden (in ruimte, tijd, of allebei) hebben waarschijnlijk soorgelijke activatieniveaus. De a-priori dichtheid houdt de regressie coëfficiënten ongecorreleerd, maar het correleert hun absolute waarden. De symmetrie eigenschappen van de a-priori dichtheden leiden tot a-posteriori dichtheden die blokdiagonale correlatiestructuren impliceren. De normaalverdeling die gebruikt wordt voor de benadering erft deze eigenschap. De blokdiagonale covariantiestructuur en de gewoonlijk ondergespecificeerde regressiemodellen zorgen ervoor dat de computationele complexiteit van EP lineair schaalt met het aantal regressiecoëfficiënten. We laten zien dat de gewichtigheidskaarten (*importance maps*) op basis van benaderde a-posteriori momenten van de schaalparameters betekenisvol en in neuro-biologische zin redelijk zijn.

Acknowledgements

The research reported in this thesis has been financially supported by the European Commission through the research grant Artificial Intelligence for Industrial Applications, a joint collaboration between SKF R&D (Nieuwegein) and Radboud University Nijmegen, and by the Netherlands Organization for Scientific Research through Tom's Vici grant.

I would like to thank Tom for giving me the opportunity to pursue a Ph.D. program and for guiding me through it. With all its ups and downs, it was a great experience and I am very happy for having the chance to work with him both on the subjects addressed in this thesis and on the others which did not make it.

I am also thankful to all former and present members of Tom's group and the former IRIS group who contributed both directly, through collaboration, or indirectly, through fruitful discussions or moral support, to the creation of the research reported in this thesis.

First of all, I would like to thank Marcel van Gerven for the cooperation that resulted in the material presented in Chapter 3. I would also like to thank Evgeni Tsivtsivadze for the research and fun we shared. It was very pleasant to work with both of you. Your enthusiasm and hard-working attitude will always be a example for me to follow. I am very thankful to Tom (C) for being a great office companion and helping me out with almost everything starting from benchmarking my "what if..."-s on the white board to giving advices in how to write job application letters. And last but not the least, for being funny and for being the most receptive and tolerant person when I was trying to do the same.

I would like to thank my former group-mates and friends Ildikó, Henriëtte, Arjen and Daan for being my tutors in the Netherlands and for helping me out whenever *babelfish* wasn't helping much, which was typically the case. I also thank for the dinners, drinks, journeys and fun we had together, especially for the great tea-times some of us had while pondering about the meaning of our lives and of the fishes in the fish-tank next to the biology department. I am also grateful to my group-mates Adriana, Perry and Rasa for the dinners, board/bored game parties we had.

I would like to thank my friends Marleen and Kaleem, as well as all the almost forty house-mates I've met during my stay, for the great time we had together at Groesbeekseweg 167. It is a great place with lots of nice people from all over the world.

Now that the thesis has been accepted for public defense, it is the time to evaluate whether I have made a good decision when embarking on the journey that ended in writing this booklet. Despite all uncertainties, which some say are natural, I would like to thank L. Csató and A. Soós for advising it to me and supporting me after I made this decision.

I would like to thank Imola, Sanyi and Márta for their remote support and for taking care of my well-being during the holidays.

Finally, I would like to thank my family for their positive attitude, their support and encouragement and for accepting and dealing with the lack of my presence whenever I could have been helpful. *Kedves édesanya, édesapa, mama és Cicus, nagyon köszönöm a sok bíztatást és segítséget.*

Nijmegen, October, 2010 Botond

SIKS Dissertatiereeks

1998:

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects
1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information
1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective
1998-4 Dennis Breuker (UM)
Memory versus Search in Games
1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

1999:

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modeling; Automated modeling of Quality Change of Agricultural Products
1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets
1999-3 Don Beal (UM)
The Nature of Minimax Search
1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects
1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Drive Specification of Network Information Systems
1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems
1999-7 David Spelt (UT)
Verification support for object database design
1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.

2000:

2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance
2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management
2000-3 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.
2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design
2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval.
2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication
2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management
2000-8 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks
2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization
2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms and Architecture
2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

2001:

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks
2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models
2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets
2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style
2001-6 Martijn van Welie (VU)
Task-based User Interface Design
2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization
2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics.
2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes
2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice BRAHMS: a multiagent modeling
and simulation language for work practice analysis and design
2001-11 Tom M. van Engers (VUA)
Knowledge Management: The Role of Mental Models in Business Systems Design

2002:

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis
2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections
2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval
2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining
2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents
2002-06 Laurens Mommers (UL)
Applied legal epistemology; Building a knowledge-based ontology of the legal domain
2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications
2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas
2002-09 Willem-Jan van den Heuvel(KUB)
Integrating Modern Business Applications with Objectified Legacy Systems
2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble
2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications
2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems
2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications
2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems
2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling
2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications
2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

2003:

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments
2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems
2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy
2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology
2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach
2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks
2003-08 Yongping Ran (UM)
Repair Based Scheduling
2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour
2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture
2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks
2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval
2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models
2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations
2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems
2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses
2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing
2003-18 Levente Kocsis (UM)
Learning Search Decisions

2004:

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic
2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business
2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving
2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures
2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity
2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques
2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes
2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise
2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning
2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects
2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies
2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents
2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play
2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium
2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining
2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning
2004-17 Mark Winands (UM)
Informed Search in Complex Games
2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models
2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval
2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

2005:

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications
2005-02 Erik van der Werf (UM))
AI techniques for the game of Go
2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language
2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data
2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing
2005-06 Pieter Spronck (UM)
Adaptive Game AI
2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems
2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications
2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages
2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments
2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search
2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry
2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen
2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics
2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes
2005-16 Joris Graaumans (UU)
Usability of XML Query Languages
2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components
2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks
2005-19 Michel van Dartel (UM)
Situated Representation
2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives
2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006:

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting
2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations
2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems
2006-04 Marta Sabou (VU)
Building Web Service Ontologies
2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines
2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling
2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering
2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web
2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion
2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems
2006-11 Joeri van Ruth (UT)

Flattening Queries over Nested Data Types
2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological environment, and the arts
2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents
2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change
2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain
2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks
2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device
2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing
2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach
2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining
2006-21 Bas van Gils (RUN)
Aptness on the Web
2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation
2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web
2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources
2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC
2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval
2006-27 Stefano Bocconi (CWI)
Vox Populi: generating video documentaries from semantically annotated media repositories
2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval

2007:

2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures
2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach
2007-03 Peter Mika (VU)
Social Networks and the Semantic Web
2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach
2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance
2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs
2007-07 Natasa Jovanovic' (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings
2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations
2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation
2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols
2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System
2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty
2007-13 Rutger Rienks (UT)
Meetings in Smart Environments; Implications of Progressing Technology
2007-14 Niek Bergboer (UM)
Context-Based Image Analysis
2007-15 Joyca Lacroix (UM)
NIM: a Situated Computational Memory Model
2007-16 Davide Grossi (UU)

Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems
2007-17 Theodore Charitos (UU)
Reasoning with Dynamic Networks in Practice
2007-18 Bart Orriens (UvT)
On the development an management of adaptive business collaborations
2007-19 David Levy (UM)
Intimate relationships with artificial partners
2007-20 Slinger Jansen (UU)
Customer Configuration Updating in a Software Supply Network
2007-21 Karianne Vermaas (UU)
Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005
2007-22 Zlatko Zlatev (UT)
Goal-oriented design of value and process models from patterns
2007-23 Peter Barna (TUE)
Specification of Application Logic in Web Information Systems
2007-24 Georgina Ramrez Camps (CWI)
Structural Features in XML Retrieval
2007-25 Joost Schalken (VU)
Empirical Investigations in Software Process Improvement

2008:

2008-01 Katalin Boer-Sorbn (EUR)
Agent-Based Simulation of Financial Markets: A modular,continuous-time approach
2008-02 Alexei Sharpanskykh (VU)
On Computer-Aided Methods for Modeling and Analysis of Organizations
2008-03 Vera Hollink (UVA)
Optimizing hierarchical menus: a usage-based approach
2008-04 Ander de Keijzer (UT)
Management of Uncertain Data - towards unattended integration
2008-05 Bela Mutschler (UT)
Modeling and simulating causal dependencies on process-aware information systems from a cost perspective
2008-06 Arjen Hommersom (RUN)
On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective
2008-07 Peter van Rosmalen (OU)
Supporting the tutor in the design and support of adaptive e-learning
2008-08 Janneke Bolt (UU)
Bayesian Networks: Aspects of Approximate Inference
2008-09 Christof van Nimwegen (UU)
The paradox of the guided user: assistance can be counter-effective
2008-10 Wauter Bosma (UT)
Discourse oriented summarization
2008-11 Vera Kartseva (VU)
Designing Controls for Network Organizations: A Value-Based Approach
2008-12 Jozsef Farkas (RUN)
A Semiotically Oriented Cognitive Model of Knowledge Representation
2008-13 Caterina Carraciolo (UVA)
Topic Driven Access to Scientific Handbooks
2008-14 Arthur van Bunningen (UT)
Context-Aware Querying; Better Answers with Less Effort
2008-15 Martijn van Otterlo (UT)
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.
2008-16 Henriette van Vugt (VU)
Embodied agents from a user's perspective
2008-17 Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises
2008-18 Guido de Croon (UM)
Adaptive Active Vision
2008-19 Henning Rode (UT)
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search
2008-20 Rex Arendsen (UVA)
Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven
2008-21 Krisztian Balog (UVA)
People Search in the Enterprise
2008-22 Henk Koning (UU)
Communication of IT-Architecture

2008-23 Stefan Visscher (UU)
Bayesian network models for the management of ventilator-associated pneumonia
2008-24 Zharko Aleksovski (VU)
Using background knowledge in ontology matching
2008-25 Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency
2008-26 Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled
2008-27 Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design
2008-28 Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks
2008-29 Dennis Reidsma (UT)
Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans
2008-30 Wouter van Atteveldt (VU)
Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content
2008-31 Loes Braun (UM)
Pro-Active Medical Information Retrieval
2008-32 Trung H. Bui (UT)
Toward Affective Dialogue Management using Partially Observable Markov Decision Processes
2008-33 Frank Terpstra (UVA)
Scientific Workflow Design; theoretical and practical issues
2008-34 Jeroen de Knijf (UU)
Studies in Frequent Tree Mining
2008-35 Ben Torben Nielsen (UvT)
Dendritic morphologies: function shapes structure

2009:

2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models
2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques
2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT
2009-04 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering
2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality
2009-06 Muhammad Subianto (UU)
Understanding Classification
2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion
2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments
2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems
2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications
2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law & the Semantic Web
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
Operating Guidelines for Services
2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems
2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies
(making ontologies work in e-science with ONTO-SOA)
2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense
2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess
2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data
2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System
2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets
2009-20 Bob van der Vecht (UU)

Adjustable Autonomy: Controling Influences on Decision Making
2009-21 Stijn Vanderlooy(UM)
Ranking and Reliable Classification
2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence
2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment
2009-24 Annerieke Heuvelink (VU)
Cognitive Models for Training Simulations
2009-25 Alex van Ballegooij (CWI)
RAM: Array Database Management through Relational Mapping
2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services
2009-27 Christian Glahn (OU)
Contextual Support of social Engagement and Reflection on the Web
2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models
2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications
2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage
2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text
2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors
2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?
2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach
2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling
2009-36 Marco Kalz (OU)
Placement Support for Learners in Learning Networks
2009-37 Hendrik Drachsler (OU)
Navigation Support for Learners in Informal Learning Networks
2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based on Petri Nets
2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language
2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings
2009-42 Toine Bogers (UvT)
Recommender Systems for Social Bookmarking
2009-43 Virginia Nunes Leal Franqueira (UT)
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients
2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations
2009-45 Jilles Vreeken (UU)
Making Pattern Mining Useful
2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

2010:

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter
2010-02 Ingo Wassink (UT)
Work flows in Life Science
2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents
2010-04 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multi-display Environments
2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems
2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI
2010-07 Wim Fikkert (UT)

Gesture interaction at a Distance
2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software.
Protecting user freedoms in a world of software communities and eGovernments
2010-09 Hugo Kielman (UL)
Politile gegevensverwerking en Privacy, Naar een effectieve waarborging
2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children
2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning
2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis
2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques
2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration
2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models
2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice
2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications
2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation
2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems
2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative
2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation
2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data
2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions
2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies
2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mind-reading Perspective
2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines
2010-27 Marten Voulon (UL)
Automatisch contracteren
2010-28 Arne Koopman (UU)
Characteristic Relational Patterns
2010-29 Stratos Idreos (CWI)
Database Cracking: Towards Auto-tuning Database Kernels
2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval
2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web
2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems
2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval
2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions
2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval
2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning;
Facilitating competence development through a learning path specification
2010-37 Niels Lohmann (TUE)
Correctness of services and their composition
2010-38 Dirk Fahland (TUE)
From Scenarios to components
2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents
2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web

2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search
2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach
2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies
2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain
2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services
2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment
2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples
2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets
2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions
2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives